AGARD-LS-199

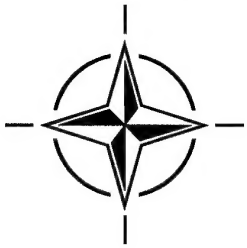
AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT
7 RUE ANCELLE, 92200 NEUILLY-SUR-SEINE, FRANCE

AGARD LECTURE SERIES 199

Optical Processing and Computing
(le Traitement optique de données et l'informatique)

This publication was prepared at the request of the Sensor and Propagation Panel and under the sponsorship of the Consultant and Exchange Programme of AGARD and will be presented on 12-13 October 1995 in Paris, France, 16-17 October 1995 in Rome, Italy, and 19-20 October 1995 in Ankara, Turkey.



NORTH ATLANTIC TREATY ORGANIZATION

Published September 1995

Distribution and Availability on Back Cover

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE, 92200 NEUILLY-SUR-SEINE, FRANCE

AGARD LECTURE SERIES 199

Optical Processing and Computing

(le Traitement optique de données et l'informatique)

This publication was prepared at the request of the Sensor and Propagation Panel and under the sponsorship of the Consultant and Exchange Programme of AGARD and will be presented on 12-13 October 1995 in Paris, France, 16-17 October 1995 in Rome, Italy, and 19-20 October 1995 in Ankara, Turkey.

North Atlantic Treaty Organization
Organisation du Traité de l'Atlantique Nord

The Mission of AGARD

According to its Charter, the mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community;
- Providing scientific and technical advice and assistance to the Military Committee in the field of aerospace research and development (with particular regard to its military application);
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Exchange of scientific and technical information;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

Published September 1995

ISBN 92-836-0018-5

Printed by Canada Communication Group
45 Sacré-Cœur Blvd., Hull (Québec), Canada K1A 0S7

Contents

Abstract

OPTICAL PROCESSING AND COMPUTING

Optical computing, namely information processing using light waves to represent the data, possesses some inherent advantage over electronic computing, in particular for massive data storage and parallel and neural processing. The main aim of the LS is to show how recent advances in lightwave technology make the time ripe to consider exploiting the potential of optical computing for data processing applications.

The LS will be opened with an overview of the basic concepts and inherent advantages of using optics for data processing and computing applications. The rest of the first day will be devoted to two topics: the use of optics for interconnecting electronic and optoelectronic processors and the use of optoelectronic techniques to enhance the performance of various computing devices and systems.

The second day of the LS will be opened with an overview of the emerging field of artificial neural networks as a signal processing paradigm, and its hardware, and in particular its optical implementations. Finally, the LS will be concluded with a description of recent developments of optoelectronic data communication, and their forecasted effect on computing and data processing technologies.

Abrégé

LE TRAITEMENT OPTIQUE DE DONNÉES ET L'INFORMATIQUE

L'informatique optique, c'est-à-dire le traitement de l'information à l'aide d'ondes lumineuses pour représenter les données, offre certains avantages naturels par rapport au calcul électronique, en particulier pour ce qui concerne la mémorisation massive de données, ainsi que le traitement parallèle et neuronal. L'objectif principal de ce cycle de conférences est de démontrer, grâce aux progrès réalisés dernièrement dans le domaine des technologies des ondes lumineuses, l'opportunité de l'exploitation du potentiel de l'informatique optique pour des applications de traitement de l'information.

Le cycle de conférences débutera par un tour d'horizon des concepts de base et des avantages inhérents à l'emploi de l'optique pour des applications de traitement de données et de calcul. Le restant du premier jour sera consacré à deux sujets: le recours à l'optique pour l'interconnexion de processeurs électroniques et optoélectroniques et la mise en œuvre de techniques optoélectroniques pour l'amélioration des performances des différents dispositifs et systèmes informatiques.

Le programme du deuxième jour s'ouvrira par un résumé du domaine embryonnaire des réseaux neuronaux artificiels en tant que paradigme du traitement du signal, ainsi que son matériel, et en particulier ses applications optiques.

Le cycle de conférences se terminera par une description des derniers développements en transmission optoélectronique de données et l'impact prévisible de cette technique sur les technologies du calcul et du traitement des données.

List of Authors/Speakers

Lecture Series Director: Professor Bruno CROSIGNANI
Dipartimento di Fisica
Universita di Roma, "La Sapienza"
Piazzale Aldo Moro 5
00185 Roma
ITALY

Authors/Speakers

Prof. Pierre CHAVEL
Institut d'Optique Théorique
CNRS
BP 147
F-91403 Orsay
FRANCE

Prof. Sadik ESENER
University of California 0407
San Diego
Dept. of Electrical & Computer Eng.
9500 Gilman Drive
La Jolla, CA 920093-5003
USA

Prof. Kerry VAHALA
Dept. of Applied Physics
California Institute of Technology
Pasadena
CA 91125
USA

Prof. Aharon AGRANAT
Department of Applied Physics
The Hebrew University of Jerusalem
Givat Ram Campus
91904 Jerusalem
ISRAEL

Co-Authors

Mr. Mike NEWKIRK
Ortel Corporation
Alhambra
California
USA

Mr. Barry MILLER
AT & T Bell Laboratories
Holmdel
New Jersey
USA

Mr. Jianhui ZHOU
AT & T Bell Laboratories
Holmdel
New Jersey
USA

Mr. Namkyoo PARK
AT & T Bell Laboratories
Murray Hill
New Jersey
USA

OPTICS AND NONLINEAR OPTICS FOR SIGNAL PROCESSING AND COMPUTING APPLICATIONS

Bruno Crosignani
Dipartimento di Fisica dell'Universita' di Roma "La Sapienza"
00185 Roma, Italy

Summary : Some basic optical concepts are presented which are at the basis of development and implementation of devices to be used in the frame of optical processing and computing.

1. Introduction

Electronics and optics deal, by and large, with handling and manipulating electrons and photons, respectively. The main potential edge of optics over electronics is that photons, unless electrons, are massless and do not practically interact among themselves. Are these properties which are basically responsible for the intrinsic parallelism of optical processing and the extremely large bandwidths achievable in the frame of optical communications. Besides, optics possesses other inherent advantages associated with the interactions taking place inside nonlinear media providing the possibility of manipulating light with light (*nonlinear optics*, NLO).

Even if integrated electronics is to date much more advanced than integrated optics (VLSI technology has the capability of accomodating about 10^{10} logic gates on a single wafer, each of them being able to perform a logic operation in a time period less than 10^{-10} sec) , optical interconnection by itself can offer many advantages. It is thus to be expected that an hybrid technology, exploiting the strenghts of both electronics and optics, will be adopted in the future as optimal for computing systems. Among the optical elements which are at the basis of optical interconnection, *holograms* play a special role since they can be tailored to act as efficient fixed interconnections between elements with different spatial geometry (see Fig. 1).

Paper presented at the AGARD SPP Lecture Series on "Optical Processing and Computing" held in Paris, France from 12-13 October 1995; Rome, Italy from 16-17 October 1995 and Ankara, Turkey from 19-20 October 1995 and published in LS-199.

The above arguments are also valid when comparing electronic and optical implementation of neural networks, whose basic architecture mimics that of biological neural systems and which consists (see Fig.2) of many identical elements (*neurons*) linked by interconnections (*synapses*). While integrated-circuit logic elements operate in nanoseconds and have dimensions of the order of microns, achieving the necessary connectivity in electronic circuits poses serious problems. They can be overcome by interconnecting neurons by means of light beams which can simultaneously propagate and overlap without interaction in three dimensions, thus going beyond the intrinsic planarity of integrated circuits. Optical implementation also requires a device capable of converting the input patterns into an appropriate format (e.g., electrical to optical) and a thresholding device for the output unit.

Despite their obvious advantages, there are many practical problems with optical implementations since optical devices have their own physical characteristic which often do not exactly match the requirements of artificial neural networks. It is thus expedient to any real understanding of the potential of their use a basic description of the principal optical processes which can be put to work to advantage in the frame of optical processing and computing.

2. **Spatial Fourier transformation property of thin lenses**

Let us consider the propagation of an input monochromatic optical beam of frequency ω from the plane z=0 to the output focal plane at z=+f through a *thin lens* L of focal length f, as depicted in Fig. 3. According to scalar diffraction theory, the input field amplitude u(x,y) at z=0 transforms at the plane 1 into

$$u_1(x_1,y_1)=\int_{-\infty}^{+\infty}dx\int_{-\infty}^{\infty}dy\,u(x,y)\,e^{-ikr}\cong\frac{ie^{-ikz}}{\lambda z}\int_{-\infty}^{+\infty}dx\int_{-\infty}^{\infty}dy\,u(x,y)\,e^{-ik[(x-x_1)^2+(y-y_1)^2]/2z}, \quad (1)$$

where we have taken advantage of the approximate relation

$$r=\sqrt{(x-x_1)^2+(x-y_1)^2+z^2}\cong z+\frac{(x-x_1)^2}{2z}+\frac{(y-y_1)^2}{2z} \quad (2)$$

and $\lambda=2\pi/k$. After the lens, the field transforms into

$$u_2(x_2,y_2)=e^{(ik/2f)(x_2^2+y_2^2)}u_1(x_2,y_2)=\frac{ie^{-ikz}}{\lambda z}\int_{-\infty}^{+\infty}dx\int_{-\infty}^{+\infty}dyu(x,y)\,e^{-i[(k/2z)(x_2-x)^2-(k/2f)x_2^2+x\,\rightarrow y]} \quad (3)$$

so that propagating it to the back focal-plane yields

$$u_3(x_3,y_3) = -\frac{e^{-ik(z+f)}}{\lambda^2 zf}\int_{-\infty}^{+\infty} dx_2 \int_{-\infty}^{+\infty} dy_2 e^{-i(k/2f)[(x_3-x_2)^2 + x \rightarrow y]}$$

$$X \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy u(x,y)\, e^{-i[(k/2z)(x_2-x)^2 - (k/2f)x_2^2 + x \rightarrow y]} \quad . \quad (4)$$

Rearranging the terms in the integrals and performing the integration over x_2 and y_2 finally yields

$$u_3(x_3,y_3) = i\frac{e^{-ik(z+f)}}{\lambda f} e^{-i(k/2f)(1-z/f)(x_3^2+y_3^2)} \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy u(x,y)\, e^{ik(xx_3/f + y\, y_3/f)} \quad . \quad (5)$$

Recalling now the definition of *double Fourier-transform* of the function u(x,y) , that is

$$\hat{u}(p,q) = \frac{1}{(2\pi)^2}\int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy e^{-ipx-iqy} u(x,y) \quad , \quad (6)$$

we can write

$$u_3(x_3,y_3) = i\frac{e^{-ik(z+f)}}{\lambda f} e^{-i(k/2f)(1-z/f)(x_3^2+y_3^2)} (2\pi)^2 \hat{u}(p=-kx_3/f, q=-kx_3/f) \quad (7)$$

or, if the input plane coincides with the front focal-plane (z=f),

$$u_3(x_3,y_3) = i\frac{e^{-2ikf}}{\lambda f} (2\pi)^2 \hat{u}(p=-kx_3/f, q=-ky_3/f) \, . \quad (8)$$

Thus, the output field is, apart from a factor, the Fourier transform of the input field u(x,y). This property is a clear manifestation of the power of *optical*

parallelism, since a complicated numerical computation, which would require an enormous number of algebraic operations (of the order of the number of points where u(x,y) is known times the number of points in the p,q plane where one wishes to evaluate the Fourier transform) is actually performed in the time the light takes to travel the optical system.

3. **Holography**

Let us consider (see Fig.4) the interference of two beams inside a *photosensitive* medium (like a photographic emulsion, where the exposure to the two beams and subsequent development gives rise to a density of silver atoms proportional to the optical intensity, that is to the square modulus of the electric field). The beam 1

$$E_1(\mathbf{r},t) = A_1(\mathbf{r})e^{i\omega t - i\mathbf{k}_1\mathbf{r}} \qquad (9)$$

contains, through $A_1(\mathbf{r})$, a spatial information and is called the *object beam* while the beam 2, called the *reference beam*, is a plane wave

$$E_2(\mathbf{r},t) = A_2 e^{i\omega t - i\mathbf{k}_2\mathbf{r}} . \quad (10)$$

The two beams interfere giving rise to an optical intensity

$$I = |E_1+E_2|^2 = |A_1|^2+|A_2|^2+A_2A_1^* e^{-i\mathbf{K}\mathbf{r}}+A_1A_2^* e^{i\mathbf{K}\mathbf{r}} \qquad (11)$$

where $\mathbf{K}=\mathbf{k}_2-\mathbf{k}_1$. The photosensitive medium undergoes a change Δn in its refractive index proportional to the intensity I, that is

$$\Delta n = n_1 |A_1|^2+n_1|A_2|^2+n_1 A_2A_1^* e^{-i\mathbf{K}\mathbf{r}}+n_1 A_1A_2^* e^{i\mathbf{K}\mathbf{r}} \qquad (12)$$

where n_1 is a constant, so that the wavefront spatial information $A_1(\mathbf{r})$ is recorded in the medium through $\Delta n(\mathbf{r})$. Since $A_1(\mathbf{r})$ typically varies on a scale large compared to 1/K , Δn is almost a periodic function of space and is called a *grating* or an *hologram.*

The complex amplitude $A_1(\mathbf{r})$ of the object beam can be recovered from the hologram by illuminating the photosensitive medium with the plane wave (*readout beam*)

$$E_3 = A_3 e^{i\omega t - i\mathbf{k}_3\mathbf{r}} , \quad (13)$$

which induces in the medium a polarization containing, among others, the term

$$P = n n_1 \varepsilon_0 [A_2 A_1^* e^{-i(\mathbf{K}+\mathbf{k}_3)\,\mathbf{r}} + A_1 A_2^* e^{i(\mathbf{K}-\mathbf{k}_3)\,\mathbf{r}}] A_3 e^{i\omega t} \quad . \quad (14)$$

If the readout beam is either co-propagating or counter-propagating relative to the reference beam ($\mathbf{k}_3 = \mathbf{k}_2$ or $\mathbf{k}_3 = -\mathbf{k}_2$, respectively) , then the polarization will radiate a field of the kind

$$E_4 = A_4 e^{i\omega t - i\mathbf{k}_4 \mathbf{r}} \quad , \quad (15)$$

where

$$\mathbf{k}_4 = \mathbf{k}_1 \text{ (i.e., } \mathbf{k}_3 - \mathbf{k}_4 = \mathbf{K} \text{ , Bragg's condition)} \quad , \quad A_4 = \chi (A_2^* A_3) A_1 \quad , \quad (16)$$

or

$$\mathbf{k}_4 = -\mathbf{k}_1 \text{ (i.e., } \mathbf{k}_4 - \mathbf{k}_3 = \mathbf{K}) \quad , \quad A_4 = \chi (A_2 A_3) A_1^* \quad , \quad (17)$$

which represents, respectively, a replica and the *phase conjugate* of the input beam.

The holographic technique can be used to storage a large number of images (object beams with different $\mathbf{k}_1$'s), each of them giving rise to its own diffraction grating: a reconstruction of each image is obtained when the hologram is illuminated with a readout beam in such a direction as to satisfy the Bragg condition with respect to the corresponding diffraction grating.

In order to distinguish between *thin* and *thick* holograms, one considers a perfectly periodic grating,

$$\Delta n = n_1 \cos \mathbf{K}\,\mathbf{r} \quad \text{for } |z| < L \quad (18)$$

and zero otherwise, where $\mathbf{K}$ is the grating wave vector and L the thickness of the grating. These quantities allows one to introduce the dimensionless parameter Q defined as

$$Q = \frac{2\pi\lambda L}{n_0 \Lambda^2} \quad , \quad (19)$$

where $\Lambda = 2\pi/|\mathbf{K}|$ is the grating period and n_0 the refractive index of the photosensitive medium. A grating is thin (*planar hologram*) if $Q<1$ while is thick (*volume hologram*) if $Q>1$. A typical planar hologram is obtained when the

photosensitive medium is a thin photographic film (as in the early days of holography) while *photorefractive crystals* (see next section) can record volume holograms. Planar holograms are completely equivalent to volume holograms as far as their capability of reconstructing the original object beam is concerned. However, they differ with respect to diffraction efficiency (which is typically low for the formers) and to their tolerance for the misalignment of the reading beam (which is poor for the latters). They also present a different *storage capacity* , defined as the maximum number of distinguishable gratings that can be stored, which is larger for volume holograms as compared with planar ones possessing the same linear dimension.

4. **Two-beam coupling in a fixed grating**

For the successive developments it is expedient to consider the propagation of two monochromatic beams in the presence of a refractive-index distribution as that associated with a spatially periodic planar (**r**=y,z) grating of the kind (see Fig.5)

$$n(\mathbf{r}) = n_0 + n_1 \cos(\mathbf{K}\,\mathbf{r} + \phi) . \quad (20)$$

If we write

$$\mathbf{E}(\mathbf{r},t) = \frac{1}{2}\mathbf{A}_1(\mathbf{r})e^{i\omega t - i\,\mathbf{k}_1\,\mathbf{r}} + \frac{1}{2}\mathbf{A}_2(\mathbf{r})e^{i\omega t - i\,\mathbf{k}_2\,\mathbf{r}} + \text{c.c.}, \quad (21)$$

with **E** orthogonal to the plane y,z , the evolution of the $A_{1,2}$'s can be worked out by solving, in the paraxial approximation, the parabolic wave equation

$$\left(\frac{\partial}{\partial z} + \frac{i}{2k}\frac{\partial^2}{\partial y^2}\right)\mathbf{E} = -i\left(\frac{k}{n_0}\right)n_1 \cos(\mathbf{K}\,\mathbf{r} + \phi)\mathbf{E}, \quad (22)$$

where $k=(\omega/c)n_0$ .If one neglects the wave diffraction and observes that cumulative exchange of power can only take efficiently place when the Bragg condition

$$\mathbf{k}_2 - \mathbf{k}_1 = \mathbf{K} \quad (23)$$

is satisfied, one obtains, after introducing Eq.(21) into Eq.(22),

$$\cos\theta \frac{dA_1}{dz} = -(\alpha/2)A_1 + i\frac{\pi n_1}{\lambda}e^{i\phi}A_2 , \quad (24)$$

$$\cos\theta \frac{dA_2}{dz} = -(\alpha/2)A_2 + i\frac{\pi n_1}{\lambda} e^{-i\phi} A_1 \quad , \quad (25)$$

where $\lambda = 2\pi/k$, 2θ is the angle between $\mathbf{k}_1$ and $\mathbf{k}_2$, and the loss term has been added phenomenologically. After setting $A_j=(I_j)^{1/2}\exp(-i\psi_j)$, Eqs.(24) and (25) become

$$\cos\theta \frac{dI_1}{dz} = -\alpha I_1 + \frac{2\pi n_1}{\lambda}\sqrt{I_1 I_2}\sin\Psi \quad , \quad (26)$$

$$\cos\theta \frac{dI_2}{dz} = -\alpha I_2 - \frac{2\pi n_1}{\lambda}\sqrt{I_1 I_2}\sin\Psi \quad , \quad (27)$$

where $\Psi = \psi_1 - \psi_2 + \phi$. Maximum power exchange occurs for $\Psi=-\pi/2$ and the solution of the set of Eqs.(26)-(27) corresponding to the boundary condition $I_1(z=0) = I_1(0)$ and $I_2(z=0) = 0$ reads in this case

$$I_1(z) = I_1(0)e^{-\alpha z}\cos^2\left(\frac{\pi n_1 z}{\lambda\cos\theta}\right) \quad , \quad (28)$$

$$I_2(z) = I_1(0)e^{-\alpha z}\sin^2\left(\frac{\pi n_1 z}{\lambda\cos\theta}\right) \quad . \quad (29)$$

5. The photorefractive effect and real-time holography

Photorefractive materials can be used, by exploting the *photorefractive effect*, as a ductile holographic recording medium for applications in the frame of optical computing and processing, the main attraction of this approach being its real-time aspect which obviates the need to develop the hologram (*real-time holography*). In this case, the grating is not fixed but it is generated in real time by the very two beams that propagates through it, a phenomenon known as *dynamic* or *real-time holography*.

The physical origin of the process is associated with the photorefractive effect (present in some crystals, like $BaTiO_3$, $LiNbO_3$, SBN ($Sr_{1-x}Ba_xNb_2O_6$), et cetera), which consists in a variation of the refractive index of the crystal proportional to the intensity pattern of two interfering beams (e.g., two plane

waves, see Fig.5). The effect takes place in impurity-doped *electrooptic* crystals (see Fig.6), in which the fixed donor impurities can be ionized by absorbing photons of the appropriate energy. The resulting electrons , excited to the conduction band, migrate away under the influence of diffusion and of any internally generated or externally applied electric field until they are captured by an empty donor (ionized either by a photon interaction or by losing its electron to a deep impurity acceptor). As a result, in steady state, the regions where the optical intensity is higher will acquire a positive charge while the dark regions will have an excess of electrons. The space-charge separation generated in this way will in turn give rise to a static space-charge field E^{sc} which will induce in the crystal, through the linear *electrooptic effect* (or Pockels' effect) an index grating $\Delta n = rE^{sc}$, where r is some *electrooptic coefficient* (see Fig.7). The refractive-index variation, for which the two interfering waves are responsible, causes, through a nonlinear self-induced mechanism, the interaction between them (*two-beam coupling*).

If we write the field as the superposition of two interfering waves of the same frequency ω,

$$\mathbf{E}(\mathbf{r},t) = \hat{\mathbf{e}}_1 A_1(\mathbf{r}) e^{i\omega t - i\mathbf{k}_1 \mathbf{r}} + \hat{\mathbf{e}}_2 A_2(\mathbf{r}) e^{i\omega t - i\mathbf{k}_2 \mathbf{r}} \quad , \quad (30)$$

then

$$\Delta n = -\frac{1}{2} n_0^3 r E^{sc} = \frac{1}{2}\left[\frac{n_1 e^{-i\phi} \hat{e}_1 \hat{e}_2 A_1 A_2^*}{|A_1|^2 + |A_2|^2} e^{-i\mathbf{K}\mathbf{r}} + \text{c.c.}\right]$$

$$\equiv \frac{1}{2}\left[\frac{n_1 e^{-i\phi} I_1}{I_0} e^{-i\mathbf{K}\mathbf{r}} + \text{c.c.}\right] \quad , \quad (31)$$

where the phase ϕ, representing the shift of the index grating with respect to the beam interference pattern, and n_1 , a real positive number, depend on K, on the material properties of the crystal and on the value of an external static field which may be applied to the crystal.

The comparison of Eq.(20) with Eq.(12) shows that photorefractive crystals can act as recording media in holography since the appropriate volume index grating is formed when illuminated by two beams of coherent light (which obviates the necessity of developing the hologram present in conventional holography).

6. **Optical four-wave mixing**

In a photorefractive material, the two co-propagating beams are coupled by the same grating they generate (see Eq.(31)) and the set of equations describing their evolution is no longer linear (recall Eqs.(24) and (25)) but becomes *nonlinear*. After setting

$$\Gamma \equiv \gamma + 2i\beta = i\left(\frac{2\pi n_1}{\lambda \cos\theta}\right) e^{-i\phi} \quad , \qquad (32)$$

it reads

$$\frac{d}{dz} A_1 = -\frac{1}{2I_0} \Gamma |A_2|^2 A_1 - (\alpha/2) A_1 \qquad (33)$$

$$\frac{d}{dz} A_2 = \frac{1}{2I_0} \Gamma^* |A_1|^2 A_2 - (\alpha/2) A_2 \, , \qquad (34)$$

where the loss factor α has been added phenomenologically. Its solution for the intensities $I_1(z)=|A_1|^2$ and $I_2(z)=|A_2|^2$,

$$I_1(z) = I_1(0) \frac{1+1/m}{1+(1/m)e^{\gamma z}} e^{-\alpha z} \qquad (35)$$

$$I_2(z) = I_2(0) \frac{1+m}{1+me^{-\gamma z}} e^{-\alpha z} \, , \qquad (36)$$

where $m=I_1(0)/I_2(0)$ is the initial intensity ratio, shows how energy can flow from beam 1 to beam 2 (or viceversa, according to the sign of γ) thus providing beam amplification whenever $\gamma > \alpha$.

A more sophisticated situation is that in which (see Fig.8) four beams are allowed to propagate simultaneously in the photorefractive medium. If two beams (let's say 2 and 3, designated as *pumps*) are much more intense than beam 1 (designated as *signal*) and beam 4, that is $|A_1|^2, |A_4|^2 << |A_2|^2, |A_3|^2$, then one can adopt the so-called undepleted-pump approximation and assume $dA_2/dz = dA_3/dz = 0$ The dynamics of the process , *four-wave mixing*, is then contained in the evolution of A_1 and A_4 which can be shown to obey the set of equations

$$\frac{dA_1}{dz} = -\frac{\Gamma}{2}\frac{|A_2|^2}{I_0}A_1 - \frac{\Gamma}{2}\frac{A_2 A_3}{I_0}A_4^* \quad (37)$$

$$\frac{dA_4^*}{dz} = -\frac{\Gamma}{2}\frac{|A_3|^2}{I_0}A_4^* - \frac{\Gamma}{2}\frac{A_2^* A_3^*}{I_0}A_1 \ , \quad (38)$$

where $I_0 = |A_2|^2 + |A_3|^2$ and $\Gamma = i\,(2\pi n_1/\lambda\cos\theta)\exp(-i\phi)$.

Since A_2 and A_3 are assumed to be constant (undepleted pump approximation) Eqs. (37) and (38) can be easily integrated and their solution, with the boundary condition $A_1(z=0)=A_1(0)$ and $A_4(z=L)=0$, reads

$$A_1(z) = \frac{e^{-(\Gamma z/2)} + q e^{-(\Gamma L/2)}}{1 + q e^{-(\Gamma L/2)}} A_1(0) \ , \quad (39)$$

$$A_4^*(z) = \left(\frac{A_3^*}{A_2}\right)\frac{e^{-(\Gamma z/2)} - e^{-(\Gamma L/2)}}{1 + q e^{-(\Gamma L/2)}} A_1(0) \ , \quad (40)$$

where $q = |A_3|^2/|A_2|^2$ is the pump intensity ratio.

The amplitude of the signal A_4 at the input face of the crystal $z=0$ is proportional to $A^*_1(0)$, which is the *phase conjugate* of the signal beam A_1, the phase-conjugate reflection coefficient ρ being given by

$$\rho = \frac{A_4(0)}{A_1^*(0)} = \frac{A_3}{A_2^*}\frac{1 - e^{-\Gamma^* L/2}}{1 + q e^{-\Gamma^* L/2}} \ . \quad (41)$$

The phase-conjugate reflectivity $|\rho|^2$ is accordingly given by

$$R = |\rho|^2 = \left|\frac{A_4(0)}{A_1(0)}\right|^2 = \left|\frac{A_3}{A_2}\right|^2 \left|\frac{\sinh(\Gamma L/4)}{\cosh(-\ln\sqrt{q} + \gamma L/2)}\right|^2 \ . \quad (42)$$

which can assume also values much larger than one.

7. **Convolution and correlation via four-wave mixing**

Spatial convolution and correlation of monochromatic fields can be obtained via four-wave mixing in the common focal plane of two lenses, as shown in Fig.9.

Fields 1 and 2 propagate in the z-direction while field 3 propagates in the -z direction. The input amplitudes $u_1(x,y)$, $u_2(x,y)$ and $u_3(x,y)$ in the outer focal planes are Fourier transformed by propagating to the common focal plane (see

Sect.2). If $\hat{u}_1(p,q)$, $\hat{u}_2(p,q)$ and $\hat{u}_3(p,q)$ are their Fourier transform, then, as a result of four-wave mixing, the phase conjugate wave is given by

$$\hat{u}_4=\chi\,\hat{u}_1^{*}\,\hat{u}_2\,\hat{u}_3 \quad , \qquad (43)$$

where χ is a constant. This wave undergoes Fourier transform and yields $u_4(x,y)$, so that, after indicating with FT the operation of Fourier transform,

$$u_4(x,y)=\chi\ FT[\hat{u}_1^{*}\,\hat{u}_2\,\hat{u}_3] \quad . \qquad (44)$$

Special cases of interest can be obtained by assuming either $u_3(x,y)=\delta(x,y)$ or $u_2(x,y)=\delta(x,y)$ (a choice corresponding to a pinhole, the associated Fourier transform being a constant).

In the first case,

$$u_4(x,y)=\chi\ 'FT[\hat{u}_1^{*}\,\hat{u}_2] \quad , \qquad (45)$$

that is

$$u_4(x,y)=\int d\xi \int d\eta u_1^{*}(\xi-x,\eta-y)u_2(\xi\ ,\eta)\ , \qquad (46)$$

so that the nonlinear optical processor is capable of performing the *cross correlation function* of u_1 and u_2.

In the second case,

$$u_4(x,y)=\chi\ 'FT[\hat{u}_2\,\hat{u}_3] \quad , \qquad (47)$$

and the processor performs the *convolution function* of u_2 and u_3, that is

$$u_4(x,y)=\int d\xi \int d\eta u_2(\xi-x,\eta-y)u_3(\xi\ ,\eta)\ . \qquad (48)$$

9. Phase conjugate Michelson interferometer and parallel image subtraction

Optical four-wave mixing provides the possibility of creating *phase conjugate mirrors* (PCM) and this opens the way to the use of a new class of interferometers in which one or more of the standard mirrors are replaced by PCM's.

Referring to Fig.10, we consider a *phase conjugate Michelson interferometer* consisting of a beam splitter BS, a regular mirror and a PCM. An incident beam of amplitude E_0 coming from the left is divided by the beam splitter into a reflected beam of amplitude rE_0 and a transmitted one of amplitude tE_0 , where r and t are the reflection and transmission coefficients, respectively. If we indicate with r' and t' the analogous quantities when the field is incident from the right side and with ρ the reflection coefficient of the phase conjugate mirror, the output amplitude of the field is given by

$$E = (r^*t + t^*r')\rho \; E_0^* \quad . \quad (49)$$

If one now recalls that, as a general consequence of time reversal symmetry,

$$(r^*t + t^*r') = 0 \; , \quad (50)$$

the output field E turns out to be *exactly zero*, independently from the path difference between the two parts of the beam. This, in turn, implies that the two contributions are of the same amplitude but π out of phase, a circumstance which can be exploited for *parallel image subtraction.*

Let us consider the geometry sketched in Fig. 11, where two transparencies, each representing an image, are inserted in each arm of a phase conjugate Michelson interferometer. If $T_1(x,y)$ and $T_2(x,y)$ are their intensity transmittances, the electric field at the output is given by

$$E_s = \rho A^*[r^*tT_1(x,y) + r't^*T_2(x,y)] \quad , \qquad (51)$$

where A is the amplitude of the incident field. After recalling Eq.(50), Eq.(51) yields

$$E_S = \rho A^* r^* t[T_1(x,y) - T_2(x,y)] \quad , \qquad (52)$$

which is proportional to the *difference* of the two images. At the other port, created by the introduction of the beam splitter BS_2, one has

$$E_A = \rho A^* [|r|^2 [T_1(x,y) + |t|^2 T_2(x,y)] \quad , \qquad (53)$$

so that, if one chooses $|t|^2 = |r|^2 = 1/2$,

$$E_A = \frac{\rho}{2} A^* [[T_1(x,y) + T_2(x,y)] \quad , \qquad (54)$$

which is proportional to the *sum* of the two images. If one of the transparencies (e.g.,1) is removed, so that $T_1(x,y)=1$,one has

$$E_S = \rho A^* r^* t [1 - T_2(x,y)] \quad , \qquad (55)$$

which represents an *inverted image* of $T_2(x,y)$.

10. **Image amplification**

The energy transfer mechanism present in photorefractive two-wave mixing, as outlined in Sect.6, can be used for image amplification. Referring to Fig.12 , the interaction takes place inside a photorefractive crystal between a strong pump beam and a weak signal beam which carries the information, the second being amplified according to Eq.(36) which is, however, valid for two plane waves. Spatial uniform amplification, which is essential for the fidelity of the process, requires the intensity gain g over the crystal length L,

$$g = \frac{1+m}{1+me^{-\gamma L}} e^{-\alpha L} \quad , \quad (56)$$

to be independent from m, which is possible only if

$$m e^{-\gamma L} >> 1 \quad , \qquad (57)$$

that is if

$$e^{\gamma L} << m \, . \qquad (58)$$

In order to deal with an image-bearing beam, one has to generalize Eqs.(35) and (36) which are strictly valid for two plane waves. If one denotes by E_p the pump-

wave amplitude and by E_m (m=1,2,...N) the amplitudes of the image-bearing waves, the intensity of the total field reads

$$I(\mathbf{r}) = I_0 + \mathrm{Re}\{E_p \sum_{m=1}^{N} E_m^* e^{-i(\mathbf{k}_p - \mathbf{k}_m)\mathbf{r}} + \sum_{q \neq m}^{N} \sum_{m=1}^{N} E_q E_m^* e^{-i(\mathbf{k}_q - \mathbf{k}_m)\mathbf{r}} \} \quad , \quad (59)$$

where $\mathbf{k}_p$ and $\mathbf{k}_m$ are the wave vectors of the pump and of the m-th amplitude of the information-carrying beam. In one neglects the interference terms between the probe waves as compared to those between the probe waves and the pump wave, then the refractive index variation associated with the photorefractive effect can be written as (see Eq.(31))

$$\Delta n = \frac{1}{2}[e^{-i\phi} \frac{n_1}{I_0} \sum_{m=1}^{N} E_m E_p^* e^{i(\mathbf{k}_p - \mathbf{k}_m)\mathbf{r}} + \mathrm{c.c.}] \quad , \quad (60)$$

where

$$I_0 = |E_p|^2 + \sum_{m=1}^{N} |E_m|^2 \quad .$$

If one assumes $\phi = \pi/2$, the set of equations describing the evolution of the intensities of the pump $I_p = |E_p|^2$ and of the probes $I_m = |E_m|^2$ reads

$$\frac{dI_p}{dz} = -\frac{\Gamma}{I_0} \sum_{m=1}^{N} I_m I_p \quad , \quad (61)$$

$$\frac{dI_m}{dz} = \frac{\Gamma}{I_0} I_m I_p \quad , \quad (62)$$

where $\Gamma_p = 2n_1\pi/\lambda\cos\theta_p$, $\Gamma_m = 2n_1\pi/\lambda\cos\theta_m$, θ_p and θ_m being the angles associated with the direction of propagation of the pump and of the probe waves. In particular, if $\cos\theta_p \approx \cos\theta_m$, so that $\Gamma_p \approx \Gamma_m = \Gamma$, the set of Eqs. (61) and (62) can be solved thus yielding

$$I_p(z) = \frac{mI_0 e^{-\Gamma z}}{1 + m e^{-\Gamma z}} \quad , \quad (63)$$

$$\frac{I_m(z)}{I_m(0)} = \frac{1+m}{1+me^{-\Gamma z}} \quad , \quad (64)$$

which implies a uniform amplification of the probe waves, the common gain factor

$$m = I_p(0) / \sum_{m=1}^{N} I_m(0) \qquad (65)$$

being a function of the ratio between the pump intensity and the total intensity of the probes.

Conclusions

We have shown how some basic processes which have been developed in the frame of linear and nonlinear optics can be used to perform functions which are relevant for optical processing and computing. The main purpose is to place into evidence how the wealth of phenomena which we are currently able to understand and control can, if properly exploited by the ingenuity of applied scientists, dramatically improve in the next years the contribution of optics to information handling and processing.

References

1. A. Yariv, "Optical Electronics", 3rd Ed. (Holt, Rinehart, and Winston, New York, 1985)

2. P. Yeh, A.E. Chiou, J. Hong, P. Beckwith, T. Chang, Monte Keshnevisan, "Photorefractive nonlinear optics and optical computing", Opt. Engineering **28**, 328 (1989)

3. P. Yeh, "Introduction to Photorefractive Nonlinear Optics", (J. Wiley, New York, 1993)

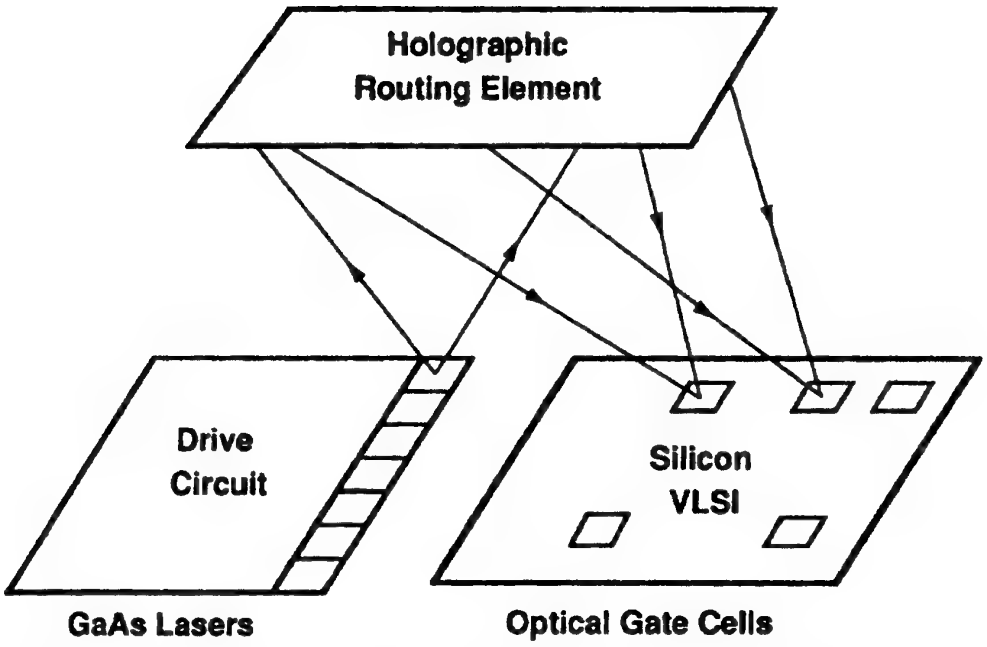


Fig.1 Typical optical interconnection geometry

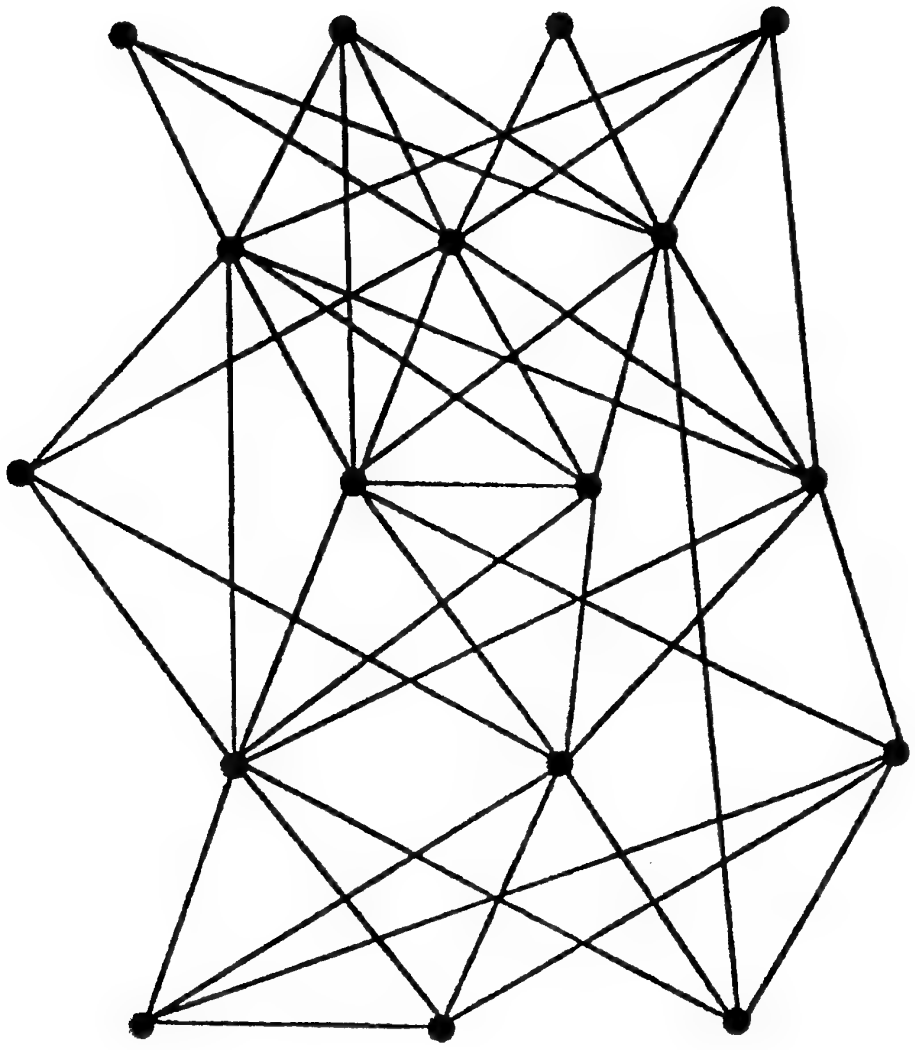


Fig.2 Architecture of neural networks

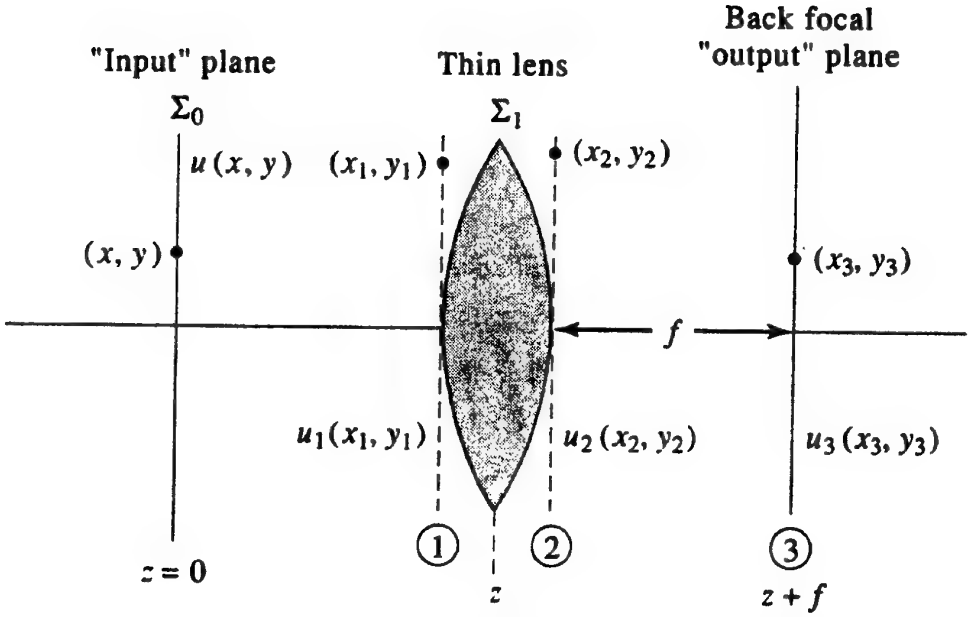


Fig.3 Fourier-transform scheme using a thin lens

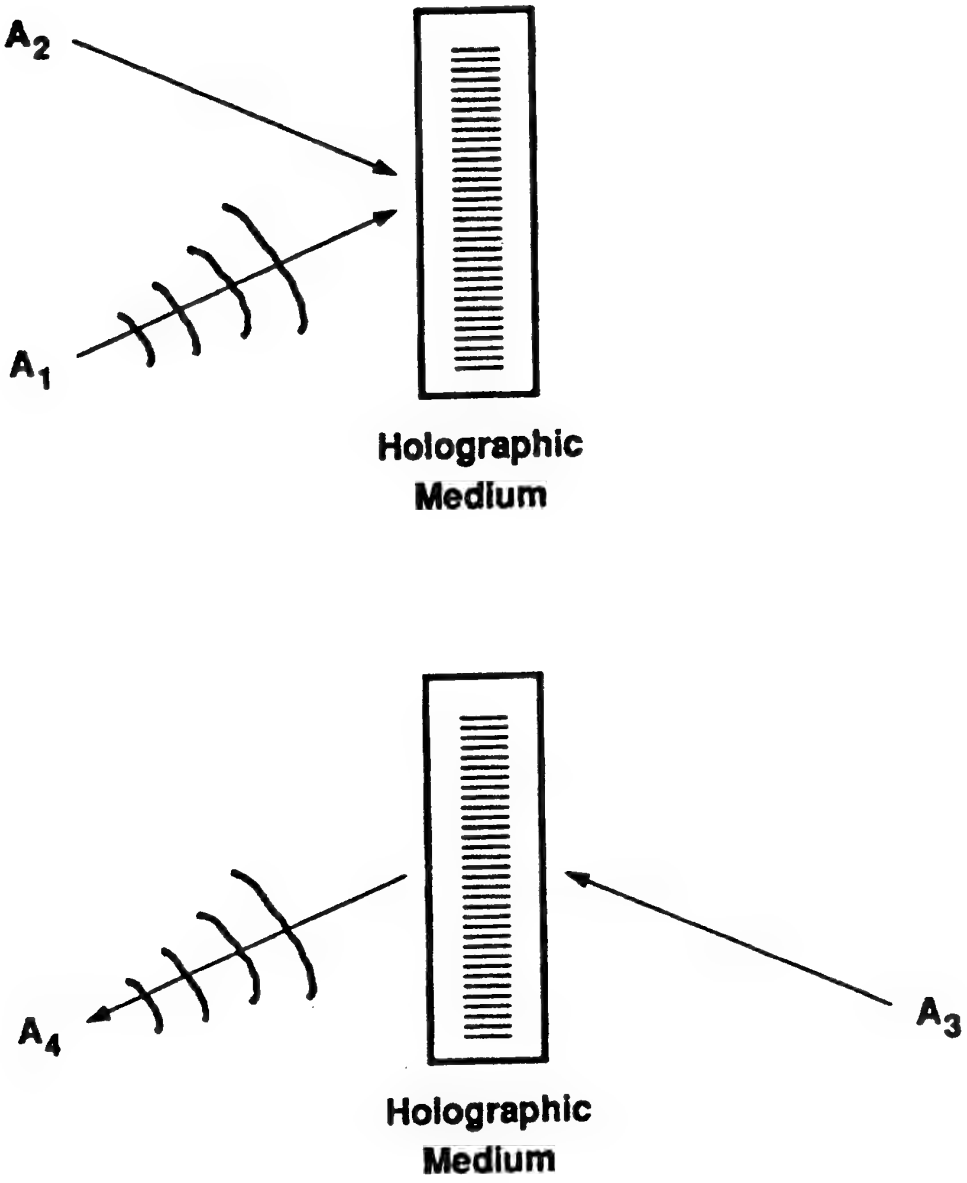


Fig.4 Holography basic scheme

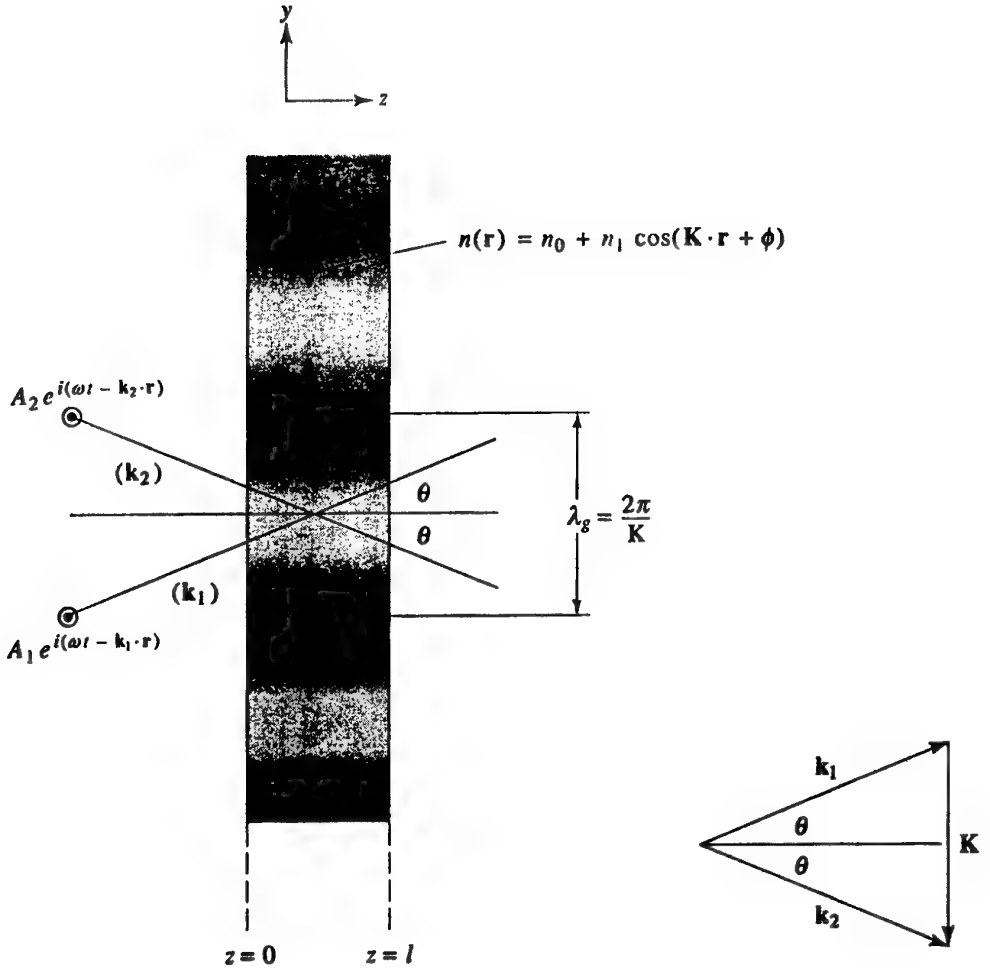


Fig.5 Two-beam coupling

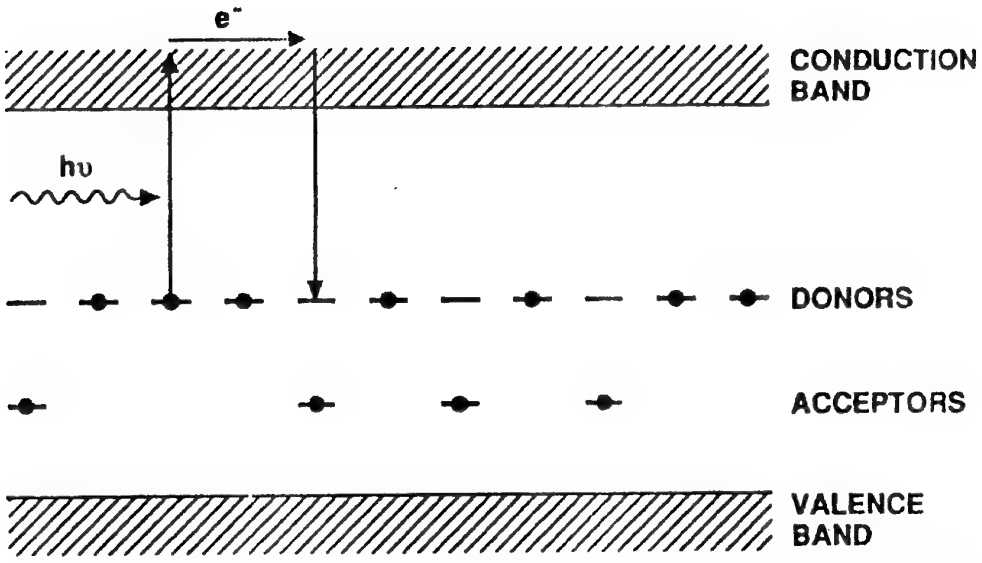


Fig.6 Charge migration and trapping in photorefractive crystals

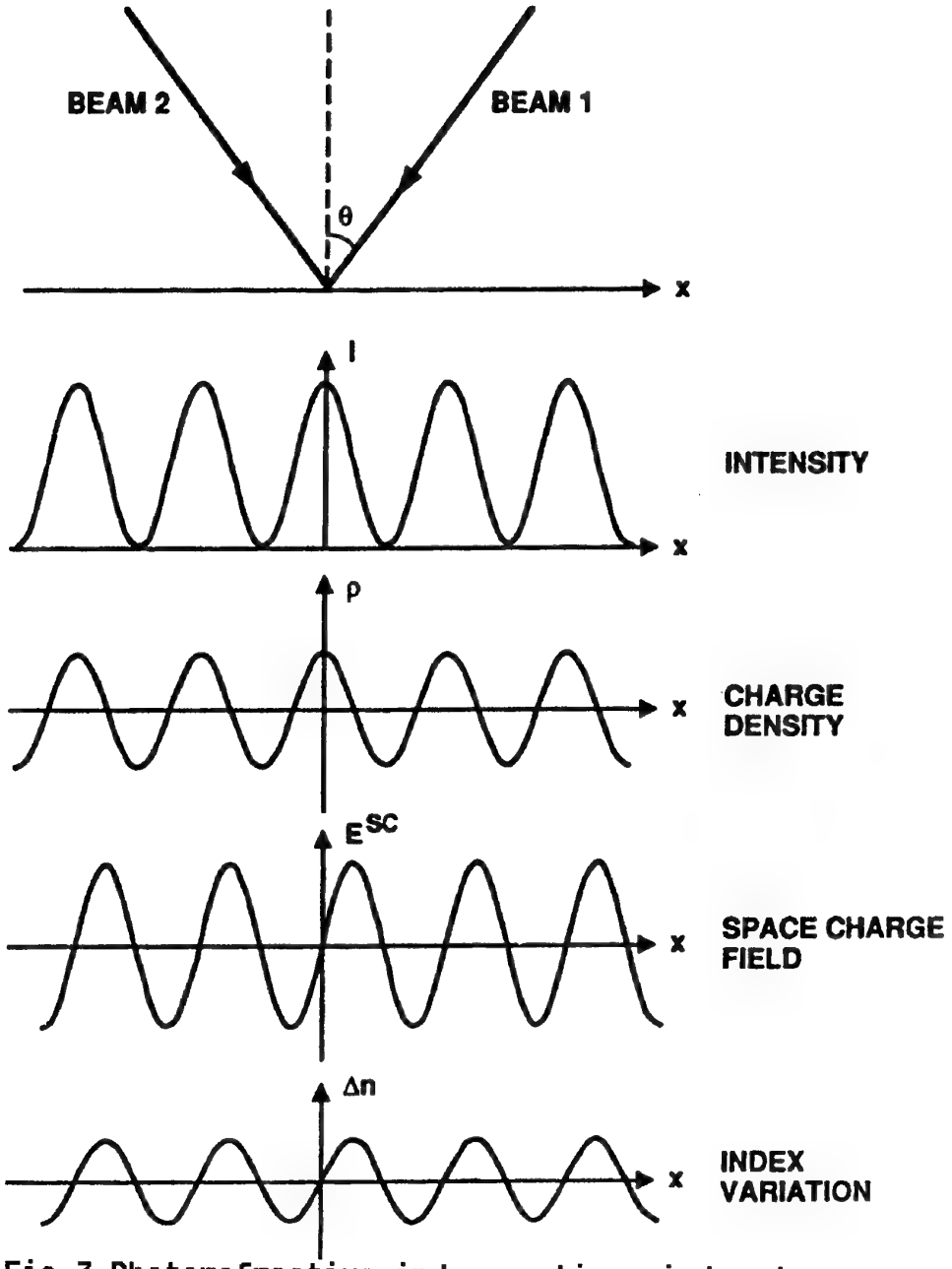


Fig.7 Photorefractive index gratings induced by two interfering plane waves

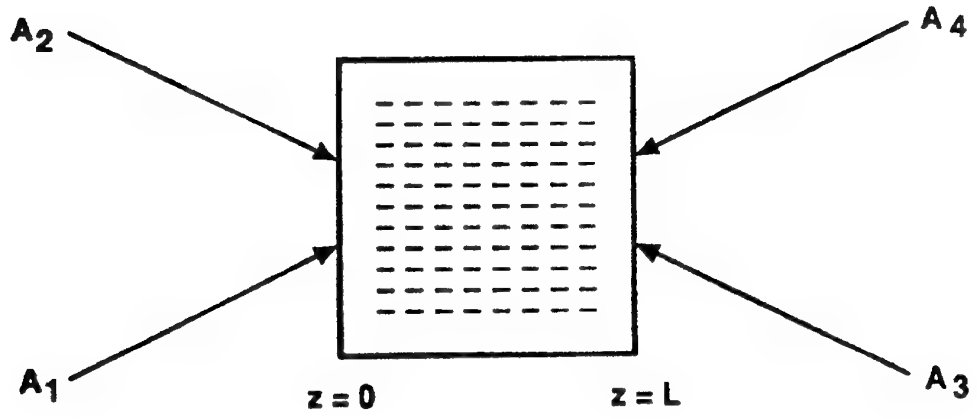


Fig.8 Four-wave mixing geometry

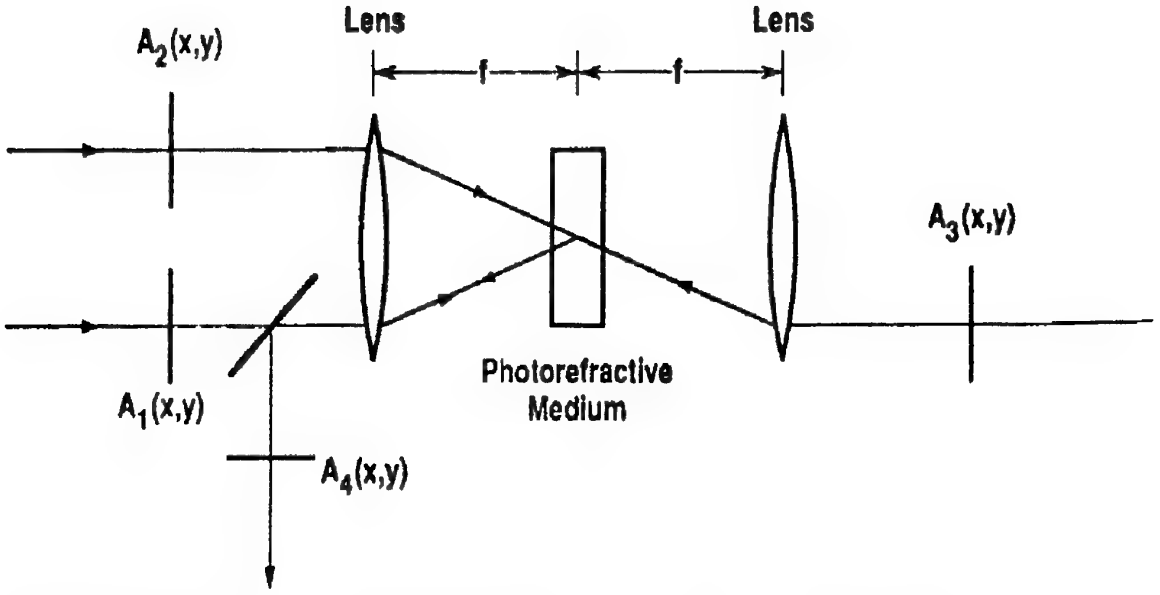


Fig.9 Convolution and correlation via four-wave mixing

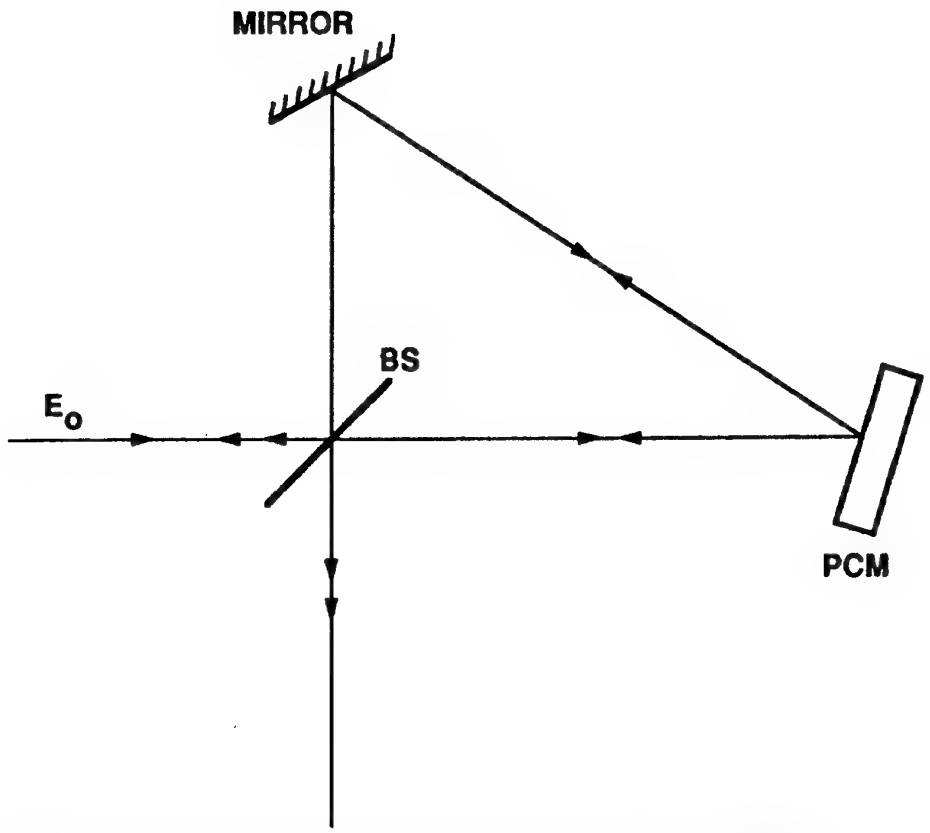


Fig.10 Phase-conjugate Michelson interferometer

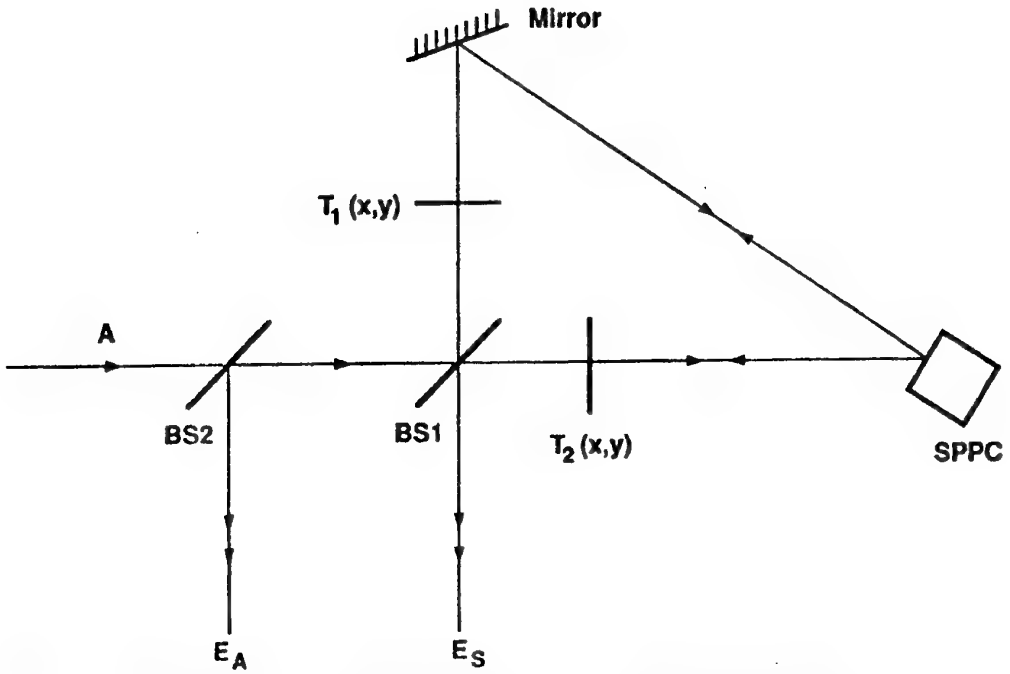


Fig.11 Parallel image subtraction scheme based on phase-conjugate Michelson interferometer

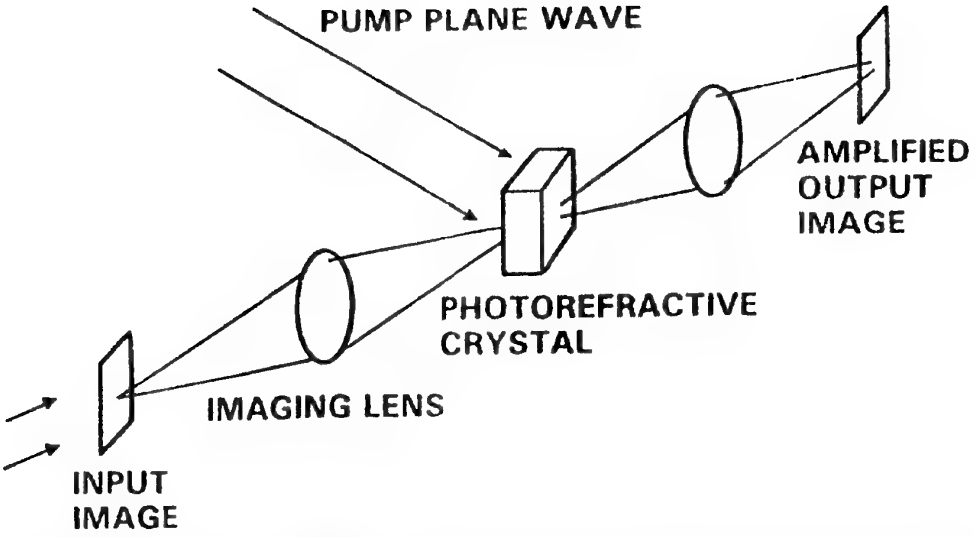


Fig.12 Optical image amplification via two-wave mixing in a photorefractive crystal

L'impact de l'optique dans les systèmes de calcul : des Interconnexions aux Machines Dédiées

Pierre Chavel
Institut d'Optique (Centre National de la Recherche Scientifique)
BP 147, 91403 Orsay cedex, France

SOMMAIRE

Tout comme dans les systèmes de télécommunications où elle est devenue omniprésente, l'optique offre des avantages dans les systèmes de calcul grâce au haut débit des communications qu'elle permet. Pour en profiter réellement, il reste à développer des techniques pour l'intégration d'un grand nombre canaux optiques dans ces systèmes et à examiner les implications sur l'architecture des machines.

Nous rappellerons tout d'abord les raisons physiques qui justifient l'intérêt porté à l'optique pour la réalisation d'interconnexions et, plus généralement, pour la définition d'architectures de machines futures. Cette analyse préliminaire explique la démarche adoptée pour la suite de ce chapitre : les technologies disponibles dictent quelques possibilités accessibles dès maintenant pour l'entrée de l'optique dans les systèmes de calcul — nous parlerons d'opto-informatique — et en laissent entrevoir de nombreuses autres. Parmi ces dernières, on peut distinguer la voie évolutionnaire et la voie révolutionnaire. La première correspond à l'addition de fonctions optiques au sein d'architectures largement dictées par la microélectronique : nous envisagerons à ce titre les réseaux d'interconnexion optiques pour machines largement parallèles. La seconde implique une révision des concepts architecturaux jusque dans le détail des circuits. Reprenant les idées de "pixels intelligents" et d'automates cellulaires, elle propose une série de processeurs dédiés entièrement nouveaux, applicables notamment à l'aide à la vision avec un parallélisme massif — nous nous expliquerons au passage sur l'emploi de ce dernier adjectif dans le contexte de l'opto-informatique.

1 - POURQUOI L'OPTIQUE ?

Nous partirons de la constatation que les fonctions à assurer dans tout système de calcul se ramènent à trois primitives : la logique binaire, la communication d'information d'un point à un autre — aussi appelée interconnexion —, et la mémorisation. Le présent chapitre ne concerne que les deux premières, la troisième étant traitée dans la contribution de S. Esener. La physique permet aisément d'identifier les arguments en faveur du recours à l'optique, ou plus exactement du mariage de l'optique et de l'électronique au profit des performances des systèmes de calcul : il s'agit de la vitesse — mais la discussion montrera qu'il est plus exact de parler en terme de délai nécessaire pour qu'une opération soit effectuée —, de la bande passante (temporelle) de communication, et de la densité (spatiale) d'interconnexion.

1.1 - "Vitesse" des fonctions optiques

On lit fréquemment que l'intérêt d'utiliser la lumière dans les ordinateurs provient de la vitesse de traitement possible. Cette affirmation lapidaire doit d'être nuancée pour éviter l'écueil de la simplification naïve. Or, tant pour les fonctions logiques que pour l'interconnexions, le temps requis par une opération n'est pas nécessairement plus court par voie optique que par voie électronique.

La logique optique utilise essentiellement l'influence d'un champ électromagnétique sur les niveaux d'énergie des solides et sur leur population. Ces phénomènes sont les mêmes que ceux de la logique électronique. Les particularités de tel ou tel matériau, de telle ou telle famille de composants peuvent justifier un léger avantage de temps de réponse au bénéfice d'un dispositif ou d'un autre, mais il ne s'agit que de facteurs assez faibles qui ne justifient en aucun cas un bouleversement des technologies et ne jouent pas systématiquement en faveur des composants optiques. Par exemple, le temps de réponse record atteint pour la commutation d'un transistor est de l'ordre de la picoseconde ; il en est de même pour la transition bistable optique la plus rapide.

Cependant, on sait que les ordinateurs actuels sont limités par la durée des communications et non pas par celle des opérations logiques, et il est vrai que la vitesse de déplacement des électrons dans les conducteurs ne dépasse pas l'ordre de grandeur du km/s alors que la vitesse de propagation de la lumière dans le vide est très voisine de 3.10^8 m/s (ou encore, en unités convenables à l'échelle d'un ordinateur, 30 cm/ns). Mais cette comparaison naïve n'a aucune pertinence : le point mérite commentaire.

En effet, la vitesse de propagation du signal n'est pas en soi un avantage pour les connexions. Le transport d'une information par voie électrique ne nécessite pas le déplacement physique d'une charge sur la même distance. Le message n'est pas transporté par les électrons, mais par le champ électromagnétique qu'ils produisent. Or ce champ est de même nature que le champ optique et se propage fondamentalement à la même vitesse : seules leurs fréquences les différencient, la fréquence des ondes optiques étant de l'ordre des centaines de térahertz

Paper presented at the AGARD SPP Lecture Series on "Optical Processing and Computing" held in Paris, France from 12-13 October 1995; Rome, Italy from 16-17 October 1995 and Ankara, Turkey from 19-20 October 1995 and published in LS-199.

(1 THz = 10^{12} Hz). Si on examine la question plus en détail, il est vrai que la vitesse à prendre en compte n'est pas celle des ondes électromagnétiques dans le vide, mais la vitesse de groupe dans le milieu considéré, qui dépend de l'indice de réfraction et de sa dispersion : dans les milieux usuels, les valeurs de l'indice à prendre en compte peuvent être légèrement en faveur de l'optique ; mais il s'agit là d'une nuance et non pas d'un avantage significatif.

1.2 - Délai de propagation et délai de communication

La discussion reste cependant incomplète tant qu'on n'a pas rappelé que le temps de propagation de l'onde électromagnétique ne s'identifie à la durée nécessaire pour communiquer une information que si le destinataire du message a mis en oeuvre les moyens de détection convenables pour percevoir le message instantanément lors de son arrivée et donc être sensible à une énergie très faible. Les compteurs de photons individuels existent aux fréquences optiques et au-delà, et les détecteurs de signaux électromagnétiques très sensibles se développent à toute fréquence. Ils sont cependant encombrants et onéreux. La question pratique dans la conception d'un ordinateur n'est donc pas de savoir si une information est parvenue, mais si la quantité d'énergie transférée au destinataire est suffisante pour atteindre le seuil de déclenchement de ses détecteurs. A l'échelle d'un système de calcul localisé dans un boîtier ou dans une baie d'électronique, le temps de propagation est faible devant le temps nécessaire pour transmettre cette énergie. Or, en pratique, la transmission d'énergie sur une ligne électrique consiste à porter une électrode à un potentiel de référence à travers une impédance donnée. Cette impédance est essentiellement due à la capacité des condensateurs parasites formés par les pistes et fils conducteurs entre eux. Sauf au voisinage immédiat de la source de lumière et du détecteur, l'utilisation d'un faisceau optique évite une grande partie de ces conducteurs et il y a donc là un gain objectif et significatif en faveur de la communication optique. Pour savoir si on peut réellement en profiter dans un cas donné, il est toutefois nécessaire d'examiner encore le bilan énergétique lié à l'émission et à la détection de la lumière et donc les technologies des sources et des récepteurs ainsi que les techniques d'intégration : nous reviendrons plus loin sur cet aspect essentiel du développement de l'opto-informatique.

1.3 - Débit d'information

Emettre une information sous forme optique, c'est moduler une source de lumière par le signal à transmettre. Cette modulation affecte la fréquence porteuse optique, dont nous venons de rappeler qu'elle est très élevée. Les télécommunications optiques utilisent de mieux en mieux la bande passante gigantesque disponible autour de cette porteuse, et la transposition de cette idée aux courtes distances caractéristiques des machines informatiques est physiquement parfaitement justifiée. Là encore, seuls le coût et l'encombrement des dispositifs performants pour le multiplexage, la détection et le décodage des informations dans une très large bande peuvent entraîner ou non la décision de recourir à l'optique.

1.4 - Densité d'interconnexion

Il existe enfin un troisième argument objectif et important en faveur des communications optiques : c'est la densité du réseau de connexions, qui est aux dimensions d'espace ce que la bande passante est au temps. L'argument intuitif et trivial que les "photons" se croisent sans interagir et que grâce à cela l'optique peut former des images riches de millions de pixels indépendants est exact. La physique permet d'envisager une densité si considérable que ses limites ultimes sont hors d'atteinte dans le contexte des réalisations technologique actuelles. La mise en oeuvre de communications optiques en grand nombre à travers "l'espace libre" est de ce fait un objectif majeur de l'opto-informatique. Elle se heurte à des problèmes de conception de systèmes et non à des limites fondamentales : l'heure est donc à l'imagination pour concevoir des solutions peu encombrantes, performantes et peu onéreuses.

1.5 - Bilan

Concluons cette analyse de façon explicite : la réduction des capacités de liaison, la bande passante disponible autour de la fréquence porteuse de la lumière et la densité potentielle des interconnexions optiques sont les justifications physiques claires au développement de l'opto-informatique. La logique optique et la possibilité de reconfiguration des réseaux optiques d'interconnexion peuvent jouer un rôle dans certains cas — on pense notamment au défi de la commutation tout optique pour les télécommunications. Mais au niveau des systèmes de calcul, elles n'interviennent qu'en second lieu. Dans l'état actuel, le développement de l'opto-informatique pour valider ces atouts est tributaire des technologies de composants et d'intégration : c'est donc à elles que sera consacré la section suivante.

2 - TECHNOLOGIES POUR L'OPTO-INFORMATIQUE

2.1 - Composants actifs

2.1.1 - Semi-conducteurs pour l'opto-informatique

Bien que d'autres familles technologiques existent, nous nous attacherons ici essentiellement sinon exclusivement aux composants à semi-conducteurs composés, dont le

développement est particulièrement rapide et la constitution particulièrement voisine de celle des composants informatiques actuels. Ce choix en partie arbitraire est imposé par la nécessité de limiter le cadre de l'exposé. L'intérêt de ces matériaux provient de la combinaison de deux facteurs : leur structure de bande et la possibilité d'étendre à ces matériaux les procédés initialement mises au point pour la micro-électronique du silicium.

On connaît la prééminence des semi-conducteurs et tout particulièrement du silicium dans les composants informatiques. Le rôle qu'y joue l'effet transistor n'a pas d'équivalent optique immédiat : en opto-informatique, une option raisonnable est de confier les opérations logiques à des transistors et de recourir à l'optique pour la transmission des données. Il faut alors nécessairement disposer de moyens d'émettre, de moduler et de détecter les faisceaux lumineux. Malgré certaines études encourageantes et bien que la détection de lumière y soit aisée, le silicium est défavorisé à ce niveau par sa structure de bande à gap indirect : sa technologie ne dispose actuellement pas de moyens très efficaces ni très développés pour réaliser l'émission et la modulation. Les semi-conducteurs composés sont donc en général mieux adaptés. Ils appartiennent notamment aux familles III-V et II-VI, ainsi en raison de la position de leurs éléments chimiques dans le tableau périodique. C'est en particulier le cas de l'arséniure de gallium, semi-conducteur à gap direct de la famille III-V qui a de bonnes capacités d'émission et de modulation et dont la technologie est de toute façon relativement développée pour des raisons indépendantes de l'optique. Associé au silicium ou indépendamment, GaAs est donc le chef de file des matériaux pour l'opto-informatique. Il peut être utilisé à l'état massif ou au contraire, au prix de moyens technologiques de pointe, sous forme d'empilements de couches d'alliages de compositions différentes associant à l'arsenic et au gallium d'autres éléments chimiques des mêmes familles : la conception de telles "hétérostructures" confère une grande souplesse pour la maîtrise des propriétés spectrales et l'optimisation des performances des dispositifs.

La physique des composants ainsi accessibles repose dans tous les cas sur l'excitation de porteurs de la bande de valence à la bande de conduction et sur l'effet d'un champ statique ou d'un éclairage sur la structure de ces bandes. Ces effets affectent les spectres d'absorption et d'émission. Nous ne les décrirons pas davantage, mais nous mentionnerons les caractéristiques qu'offrent à l'utilisateur quelques composants actuels.

2.1.2 - Lasers à semi-conducteurs

Un des composants optoélectroniques les plus courants et les plus indispensables, développé initialement pour ses applications dans d'autres domaines, est bien sûr la diode laser. La structure la plus habituelle, rappelée sur la Figure 1, utilise une longueur de diode relativement grande, de l'ordre de 100 µm à 1 mm, pour permettre l'amplification du faisceau à émettre. Des progrès récents sur la conception et la croissance des empilements de couches semiconductrices ont permis d'atteindre l'effet laser sur des dimensions nettement plus petites et assurer de ce fait l'amplification et l'émission "verticale" de lumière — le mot vertical est utilisé ici pour désigner la direction perpendiculaire au substrat. La Figure 2 présente l'allure typique des lasers à cavité verticale ou VCSEL (vertical cavity, surface emitting lasers), dont la structure se prête à la fabrication sous forme de matrice.

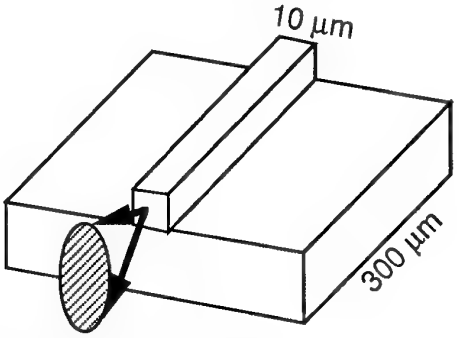


Figure 1. Laser à cavité "horizontale".

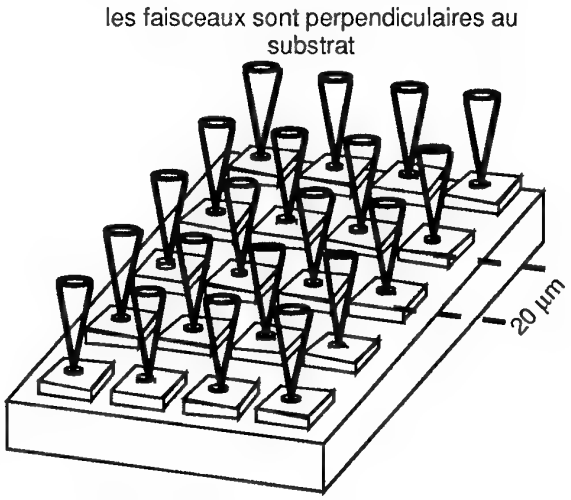


Figure 2. Matrice de lasers à cavité verticale.

L'annonce de la première matrice de lasers par Jewell en 1989 [1] a ouvert de grands espoirs. Pour la prouesse de la démonstration, la première puce contenait 2 millions de lasers par centimètre carré, c'est à dire une densité comparable à celle atteinte pour les transistors. Cependant, il est intéressant de remarquer que leur adressage individuel par le même nombre de paires d'électrodes est inaccessible à la technologie : un plan de masse commun était inclus dans la structure, mais le fonctionnement de ces lasers de démonstration nécessitait

le contact individuel de chaque élément par une pointe de test pour fournir le potentiel d'alimentation : seule une interconnexion optique serait donc envisageable pour l'excitation séparée d'un si grand nombre de voies. Depuis cette première, des matrices de lasers à adressage individuel ont été développées, elles atteignent quelques dizaines d'éléments et sont disponibles commercialement en petites quantités. On trouvera quelques exemples de performances en référence 2.

2.1.3 - Modulateurs

Les faisceaux lasers peuvent être modulés soit par le courant d'excitation de la source soit par un modulateur externe. Ce dernier mode est particulièrement adapté pour atteindre des fréquences très élevées, les records actuels étant de quelques dizaines de gigahertz [3] pour les applications aux télécommunications. Dores et déjà, le gigahertz est atteint pour les liaisons commerciales à grande distance. La même solution peut-elle être appliquée aux connexions à haut débit dans les systèmes de calcul ? La mise en oeuvre de telles bandes passantes par combinaison de signaux occupant des bandes passantes plus faibles implique les moyens de multiplexage et de démultiplexage dont l'encombrement et le coût ne peuvent être ignorés dans la conception d'un système.

Le débat entre émission de lumière sur la puce et modulation électro-optique d'un faisceau d'éclairage par la puce reste ouvert. La modulation, bien sûr, nécessite l'émission de lumière à l'extérieur de la puce par un laser qui constitue l'équivalent optique de l'alimentation des circuits électriques. Elle présente en contrepartie l'avantage d'une consommation de puissance moindre au niveau de la puce modulatrice. Le modulateur a grossièrement la même structure électrique de diode qu'un laser, mais est utilisé en polarisation inverse et donc sous haute impédance. On trouvera un exemple de performances avec l'application décrite en référence 4.

2.1.4 - Bistables

A un niveau de fonctionnalité plus élevé que le modulateur, on trouve l'élément logique tout optique. Souvent baptisé par abus de langage "transistor optique", il a en commun avec le transistor utilisé comme un élément logique la propriété de posséder deux états bien définis et aisément discernables, mais le phénomène de conduction à travers une jonction polarisée en inverse appelé "effet transistor" n'y intervient pas. Un faisceau de commande de puissance P y est absorbé et module la transmission T ou la réflexion R du dispositif. On peut donc le considérer comme un commutateur de lumière, un "volet" à commande optique. Le cas le plus souvent cité est sans doute le bistable optique, dont la caractéristique R(P) présente une boucle analogue au cycle d'hysteresis des matériaux magnétiques. La référence 5 décrit le record actuel de seuil de bistabilité en éclairage continu.

2.1.5 - Circuits logiques optoélectroniques intégrés

Décrivons plus spécialement deux composants particuliers qui peuvent donner lieu à la réalisation de circuits : le SEED et le photothyristor optique PnpN.

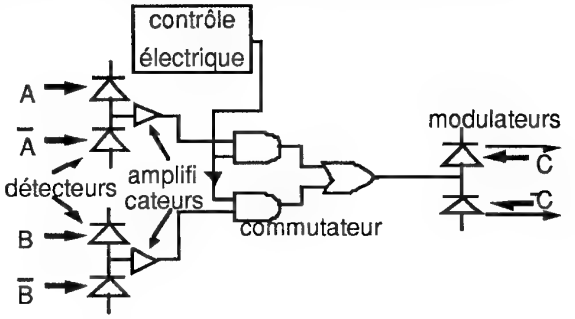


Figure 3. Exemple de circuit optoélectronique intégré réalisable à base de FET-SEED : ce circuit comprend six SEED disposés en trois paires, deux utilisées en entrée optique et une en sortie optique de données.

Le SEED (self-electro-optic effect device) a été introduit par Miller aux laboratoires ATT Bell en 1985 [6]. Sa structure a été conçue de façon telle que l'application d'une tension déplace par effet électro-optique la largeur de bande interdite et donc la transmission spectrale du composant, qui, éclairé sous la longueur d'onde convenable, devient ainsi opaque sous éclairage. Associés, sous le nom de FET-SEED, à un transistor à effet de champ chargé d'amplifier le photocourant produit par l'éclairage, ainsi que le schématise la Figure 3 dans le cas d'une cellule de commutation complexe, les SEEDs présentent actuellement les caractéristiques suivantes : énergie d'activation optique 20 à 30 fJ, cadence 650 Gbits/s (ce qui est supérieur au rythme maximal de tout circuit) [7]. ATT a ouvert leur utilisation à des collaborations externes pour réaliser des démonstrateurs (mais non pas pour l'instant des systèmes compétitifs sur le plan commercial). L'association de fonctions micro-électronique sur arséniure de gallium avec les modulateurs bistables de la famille SEED constitue un exemple majeur du développement actuel des réseaux de "smart pixels", cellules logiques dont "l'intelligence" réside en fait dans l'existence d'une ou plusieurs entrées et sorties optiques sur la surface du circuit — à moins qu'on ne les considère comme des détecteurs et modulateurs optoélectroniques dont "l'intelligence" réside dans l'existence d'un circuit logique associé à chacun d'eux.

En 1988, deux équipes ont simultanément proposé l'association d'une structure de thyristor à l'émission de lumière [8]. Il en est résulté notamment le photothyristor optique PnpN, développé par IMEC en Belgique, et qui comporte à la fois une entrée optique,

une entrée électronique et une sortie optique. Comme dans le SEED, l'entrée optique se fait sous forme de photosensibilité, en l'occurrence dans la région de la gâchette qui commande le photothyristor, l'entrée électrique se fait par la tension de la même gâchette, mais la sortie revêt dans ce cas la forme d'une diode électroluminescente constituée par le thyristor lui-même à l'état passant. Associé en paires d'éléments placés en parallèle, le tout en série avec une résistance de charge commune, les PnpN constituent un amplificateur différentiel à seuil intéressant puisque l'application de la tension de commande permet de déclencher uniquement celui des deux éléments de la paire qui a reçu le plus de lumière. Le seuil optique de photodétection différentielle atteint dans ce cas un record de sensibilité [9] pour un élément logique optique intégré avec environ 20 fJ/μm^2 (ce chiffre n'inclut pas l'énergie électrique nécessaire à la commande).

2.2 - Micro-optique :

2.2.1 - le défi de l'intégration opto-électronique

Grâce à des composants tels que ceux évoqués dans le paragraphe précédent, on peut dire que la panoplie des composants actifs pour les fonctions optoélectroniques est maintenant bien développée, au moins au niveau des laboratoires. Leur mise en oeuvre au service des systèmes informatiques requiert maintenant deux séries d'efforts complémentaires : des études de systèmes qui permettront de préciser leur apport pratique et des études d'intégration de ces systèmes qui assureront leur réalisation pratique dans des conditions fiables et crédibles. Avant de citer des exemples de systèmes actuels, commençons par examiner ce dernier aspect.

Un argument souvent cité et exact au détriment du développement de l'opto-informatique est la nécessité d'aligner les composants avec une précision de l'ordre de la dimension des composants. Augmenter la densité, diminuer la consommation, c'est nécessairement diminuer les dimensions. La connexion optique entre dispositifs implique évidemment une mise en place précise : si un élément SEED est doté, par exemple, de fenêtres optiques carrées de 10 µm, il faut bien que le faisceau qui les éclaire soit aligné à mieux que 10 µm près. La solution à ce même problème retenue pour les systèmes électroniques consiste à distinguer plusieurs niveaux : sur les puces, la précision est garantie par la lithographie ; dans les circuits hybrides, l'assemblage, délicat, est effectué une fois pour toutes ; dans les baies électroniques, l'enfichage des cartes en fond de panier, en général au pas de 2,54 mm, est défini avec une précision modeste avec des tolérances de plusieurs dixièmes de millimètres : la considérable diminution de compacité mise en évidence par ces valeurs est rendue nécessaire par les contraintes de l'assemblage. Les mêmes contraintes pratiques s'appliquent à l'optique. Pour profiter de la densité potentielle des liaisons optiques, il est donc indispensable de mettre au point des méthodes d'hybridation, d'alignement et d'intégration monolithique des différentes fonctions de manipulation de faisceau : tel est l'objet du développement actuel de la micro-optique appliquée à l'opto-informatique.

Ces fonctions sont essentiellement de trois types :
- le changement de direction, assuré par des miroirs ou des prismes ;
- la focalisation, assurée par des lentilles ou des miroirs sphériques ;
- la division de faisceau, assurée par des lames séparatrices.

Elles sont assurées, bien sûr, à l'aide des trois fonctions habituelles des composants optiques passifs : la réfraction, la réflexion, la diffraction.

2.2.2 - Composants micro-optiques :

On sait que la réfraction et la réflexion, décrites par les lois de Descartes, permettent de commander la direction et convergence d'un faisceau à l'aide de la répartition spatiale des indices de réfraction, le plus souvent à l'aide de dioptres plans ou sphériques. La fabrication de lentilles, de miroirs et de prismes est en général individuelle et nécessite des étapes de découpe et de polissage à l'abrasif. La fabrication de matrices de tels éléments interdit le polissage individuel et nécessite donc le recours à des procédés nouveaux, comme le gonflement de matériaux vitreux par échange d'ions ou de polymères par diffusion de réactifs [10]. Le contrôle précis des formes par ces méthodes pose à l'opticien des défis technologiques nouveaux dont dépend la qualité des faisceaux formés par les éléments micro-optiques, et donc, par exemple, la qualité de la focalisation sur un ensemble de récepteurs ou de bistables optiques intégrés.

Rappelons que la diffraction intervient toujours pour fixer une limite inférieure à la dimension des taches de focalisation : si un faisceau lumineux longueur d'onde λ a dans un plan (P) une section de diamètre D, sa direction ne peut pas être définie à mieux que λ/D près. Il en résulte qu'une lentille qui le fait converger à la distance f du plan (P) ne peut en aucun cas réduire son diamètre à moins de $\lambda f/D$, ou plus exactement à moins d'une valeur $V(\lambda,D,f)$ dont la forme asymptotique, rapidement atteinte, pour $D>>f$ est $\lambda f/D$ alors que la forme asymptotique pour $f<<D$ est λ. Mais la diffraction n'est pas uniquement une limitation. On peut aussi volontairement structurer une surface en éléments de petite taille D de façon que la combinaison (interférentielle) entre les faisceaux diffractés par les différents éléments ait un comportement donné. On gagne ainsi une souplesse nouvelle pour la mise en forme des fronts d'onde lumineux. Par exemple, on peut fabriquer des lentilles planaires, ou encore diviser un faisceau incident en trois faisceaux émergents ou

davantage avec une répartition contrôlée d'éclairement : la Figure 4 présente l'éclairage uniforme de 16 points grâce à un réseau de profil adapté, appelé réseau de Dammann [11]. Tel est le domaine de l'optique dite diffractive, encore connue sous deux autres noms : une forte analogie avec le principe de l'holographie justifie l'emploi de ce mot comme synonyme ; par ailleurs, la fabrication de composants optiques diffractifs profite depuis quelques années des techniques de lithographie développées pour la microélectronique : la gravure des éléments couche par couche a suggéré l'expression d'optique binaire, quelque peu trompeuse dans ce contexte.

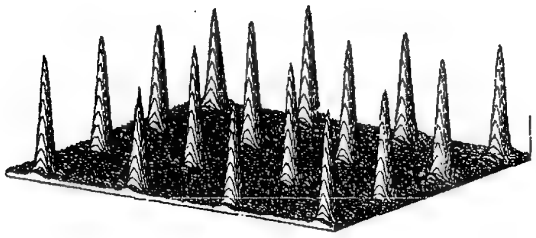
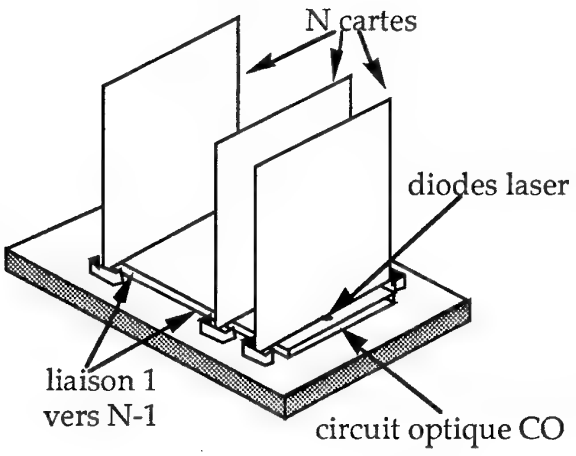


Figure 4 Distribution d'éclairement créée par un répartiteur holographique à 16 points.

3 - INTERCONNEXIONS OPTIQUES : QUELQUES EXEMPLES

Nous décrivons dans ce paragraphe trois études actuelles. La première, caractéristique des travaux entrepris par Thomson CSF dans le cadre des projets ESPRIT successifs Olives et Holics, est proche de la réalité industrielle. Les deux autres, plus en amont, résultent de travaux en cours à l'Institut d'Optique en collaboration avec différents autres laboratoires.

3.1 - Fond de panier optique



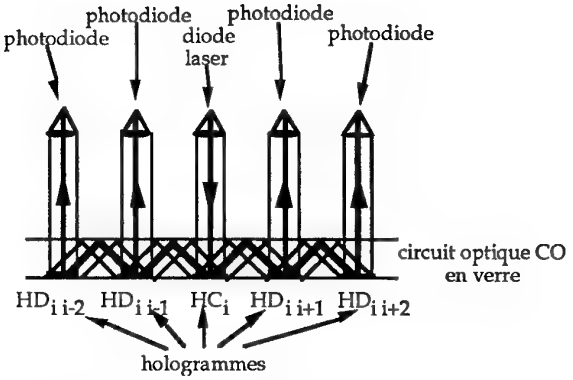


Figure 5 Connecteur optique de fond de panier pour baie électronique (en haut, perspective ; en bas, coupe).

L'idée de ce projet [12] est d'augmenter le nombre de canaux de communication entre cartes dans une baie d'électronique constituée, typiquement, de N cartes (N de l'ordre de 10) enfichées dans un connecteur de fond de panier, les différents connecteurs étant reliés entre eux par un circuit placé perpendiculairement aux cartes (voir Figure 5). Chaque connecteur peut contenir typiquement 200 broches électriques limitées par les problèmes d'intermodulation à environ 100 MHz, il s'agit d'augmenter la densité des connexions en rajoutant pour chaque carte une entrée et une sortie de connexion optique et entre les cartes un circuit optique (CO) qui leur est perpendiculaire : on garde donc exactement la géométrie habituelle de l'architecture de montage électronique, et on lui adjoint des connexions optiques. La sortie optique de chaque carte est constitué d'un laser et l'entrée comprend une série de N-1 photodétecteurs, tous alignés en nez de carte. A chaque laser et à chaque photodétecteur est associé un hologramme. La lumière du laser de la carte i, i = 1 à N, atteint immédiatement un hologramme de couplage HC_i sur le circuit CO. Le circuit est réalisé sur un substrat de verre et la lumière est envoyée par l'hologramme à l'intérieur du substrat où elle est guidée en réflexion totale (comme dans une fibre optique ou une fontaine lumineuse) jusqu'à un hologramme de découplage HD_{ij} situé face à l'un des photodétecteurs de la carte j=i+1 ou de la carte j=i-1. Une partie de la lumière est extraite de la carte par HD_{ij} pour être envoyée sur le détecteur correspondant, le reste se propage de la même façon jusqu'à la carte suivante. A l'aide de N hologrammes de couplage HC et de N(N-1) hologrammes de découplage HD, on arrive ainsi à établir un réseau complet bidirectionnel entre les cartes. Le débit des interconnexions ainsi réalisées ne dépend que de la bande passante de modulation du laser, c'est à dire en pratique de la complexité du module électronique de pilotage.

3.2 - Commutation tout optique par reconnaissance d'adresse

Cette expérience, décrite en référence 4, constitue le démonstrateur final du projet pilote du Ministère Français chargé de la Recherche sur les Matrices Optoélectroniques pour le Traitement du Signal (1987-1994). Le but est d'illustrer la possibilité de traitement parallèle avec les nouveaux composants optiques non linéaires et optoélectroniques développés par les partenaires lors des phases précédentes du projet. Le démonstrateur est un décodeur d'adresse à 64 canaux, dans lequel le chemin du signal optique d'entrée est déterminé par une adresse binaire codée séquentiellement dans le faisceau d'entrée lui-même en tête de message(voir Figure 6). La réalisation optique utilise deux composants actifs essentiels : une matrice de 8x8 modulateurs électro-optiques à multiples puits quantiques (en anglais MQW) de Thomson LCR et une matrice bistable optiques fabriquée par le CNET (laboratoire de Bagneux). Elle pourra servir de point de départ à des développements aussi bien sur les télécommunications que sur l'interconnexion des systèmes informatiques.

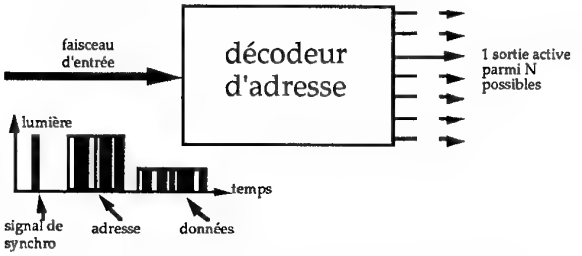


Figure 6 Fonction de reconnaissance de l'adresse d'un paquet de données.

Le principe de fonctionnement du démonstrateur est le suivant voir Figure 7) : on initialise les éléments bistables dans leur état de haute réflectivité par 64 faisceaux de maintien égaux générés à partir d'un même laser par un hologramme répartiteur de Dammann (voir plus haut). Le faisceau de signal, divisé de même en 64 parties égales, éclaire chaque pixel de la matrice de modulateurs électro-optiques. Ainsi tous les 64 modulateurs élémentaires reçoivent la même séquence lumineuse : une impulsion de synchronisation suivie de l'adresse binaire du canal de destination en codage complémenté (le bit "0" est codé par un paire d'impulsions état bas - état haut, le bit "1" par un paire haut - bas) puis par le message binaire à transmettre à travers de la voie sélectionnée. A la réception du signal de synchronisation, chacun des 64 modulateurs élémentaires émet une séquence de 12 impulsions qui représente son adresse en codage complémenté inversé. Ainsi, parmi les 64 faisceaux qui traversent le modulateur spatial, un seul ne dépasse jamais le seuil de commutation des éléments bistables, c'est celui de la voie de destination. Les 63 autres bistables passent à l'état de basse réflectivité. Le message codé en niveaux inférieurs à l'intensité de seuil n'est transmis que par le canal sélectionné.

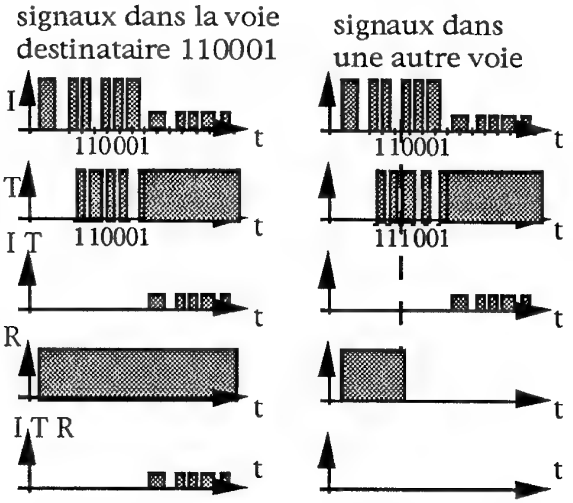


Figure 7 Principe de la logique de reconnaissance d'adresse par voie optique. I = lumière incidente. T = transmission du modulateur électro-optique. R = réflectivité du bistable.

Dans l'état actuel de l'expérience, la reconnaissance est effectuée à la cadence de 10 MHz (donc en 600 ns pour l'adresse entière), mais elle est limitée par nos moyens électroniques et non par les performances des composants opto-électroniques et optiques non linéaires. Toutefois, l'uniformité de surface et l'appariement de ces derniers limite actuellement le rapport d'extinction des faisceaux.

3.3 - Commutation tout optique par redirection de faisceau

Au cours d'une récente étude entreprise en commun avec l'INRIA, l'Université de Jérusalem et SupElec Metz, nous avons illustré expérimentalement une opération de redirection de faisceau. L'enjeu à terme est de réaliser un coffret de commutation optique pour la communication entre les baies d'un système multiprocesseur réparti sur N baies, avec N de l'ordre de quelques dizaines.

Les performances actuelles, déterminées par le financement disponible, se limitent à la possibilité d'envoyer un signal, à la cadence de 100 Mbits/s, d'un site A soit vers un site B soit vers un site C soit à la fois vers B et C. Les choix technologiques étant déterminés par la disponibilité de composants commerciaux, on recourt à l'effet acousto-optique. Un cristal acousto-optique est un dispositif capable de défléchir un faisceau lumineux en fonction d'un signal de commande acoustique, qui lui-même est issu d'une commande électrique grâce à un transducteur piézoélectrique. On peut envisager une déflexion globale de tout le faisceau par le cristal aussi bien que sa séparation en plusieurs faisceaux de sortie dont les

directions sont choisies à l'intérieur d'un domaine accessible de quelques degrés.

La perspective à terme est nettement plus ambitieuse. Elle concerne un coffret optique doté de N entrées de données, N entrées d'adresses et N sorties de données destiné à fournir une connexion complète et reprogrammable entre N sites interlocuteurs. Chaque entrée de données et chaque sortie de données est elle-même constituée de M fibres optiques qui transportent, typiquement, les M bits d'un mot dans un format convenable. Chaque entrée d'adresse peut recevoir un code de N bits provenant d'un des N sites interlocuteurs, les bits à 1 désignant les destinataires. Lorsque la baie i, i = 1 à N, doit transmettre un signal, elle envoie tout d'abord par voie électrique ce code de N bits à son entrée d'adresse de façon à identifier les baies de destination du message. Au bout d'un temps τ de l'ordre de 1 microseconde, limité par la vitesse de propagation des ondes acoustiques dans les cristaux acousto-optiques, l'adresse est en place et la baie i envoie les données : pour cela, M lasers modulés à la cadence convenable en mode bits parallèles éclairent les M fibres du canal d'entrée de données de ce même site i dans le coffret de connexion. A la sortie, et avec un délai déterminé par la longueur à parcourir, chacun des sites destinataires reçoit l'émission par les M fibres du canal de sortie du coffret qui le concerne.

4 - LES CLASSES DE PROCESSEURS OPTOELECTRONIQUES :

4.1 - Le niveau de parallélisme de l'optique :

En dehors des interconnexion, un domaine où des solutions concurrentielles pourraient venir de l'optoélectronique est le traitement d'images en milieu naturel. La raison essentielle de cette affirmation, tout comme pour les interconnexions optiques dans les systèmes électroniques, est le nombre d'opérations d'interconnexions requis pour réaliser une tâche de traitement significative : ce nombre est particulièrement grand dans les processeurs d'images parallèles, et c'est donc dans ce domaine que l'optique offre un attrait.

Le mot "interconnexion" doit être compris ici en un sens plus large que dans les paragraphes précédents : il comprend l'entrée et la sortie des données aussi bien que la fourniture de signaux de contrôle aux processeurs élémentaires (PE) de la machine parallèle et que la communication de données entre PE. Nous nous intéresserons ici au "niveau de parallélisme optique", dans lequel la machine comprend un nombre de PE égal au nombre de pixels dans l'image, c'est à dire 10^4 à tout le moins. Cette configuration exige un nombre de communication particulièrement élevé mais facilite considérablement le contrôle du programme et la structure spatiale des transferts de données. Pour qu'il soit aisé d'entrée des données sous forme d'images, tous les PE doivent tenir sur une puce optoélectronique de quelques centimètres carrés. Mais sur une si petite surface, la technologie ne permet d'envisager l'intégration que de PE extrêmement frustes. Dans ce contexte, les chercheurs se trouvent confrontés à un double défi :
- concevoir des architectures optoélectroniques qui profitent au mieux des interconnexions optiques pour maximiser la puissance de calcul,
- et trouver des algorithmes qui peuvent les utiliser pour des tâches de traitement d'image réellement utiles.
Certes, cette approche ne se situe pas bien dans le cadre des recherches actuelles sur l'architecture et la conception des systèmes informatiques parallèles. C'est qu'elle vise au contraire le développement de systèmes de traitement tout différents, très spécialisés dans certaines tâches de traitement d'images mais très puissants dans ce contexte. Nous introduisons notre point de vue à partir de l'architecture bien connu des corrélateurs (ou convolueurs) optiques et du concept d'automate cellulaire optique.

4.2 - De la convolution optique à l'automate cellulaire optique

Le cas le plus simple et le plus connu de système de traitement entrant dans cette catégorie est le convolueur optique, où la partie électronique des PE se limite aux photocapteurs et où les interconnexions optiques (analogiques) qui constituent la réponse percussionnelle réalisent en fait toute le traitement. L'application la plus évidente est la reconnaissance des formes. Le double défi dont il était question plus haut se ramène alors à ce cas bien connu : la convolution peut-elle aider à résoudre des cas réels de reconnaissance des formes, et si oui les implantations optiques sont-elles adaptées ? La reprogrammation et l'adaptation des filtres pour tenir compte des propriétés de l'image à traiter et des invariances recherchées ont enregistré ces dernières années des progrès considérables [13]. Par ailleurs, des montages de convolution optique compacts et robustes ont été présentés[14].

Néanmoins, la convolution n'est pas une opération d'une grande généralité et il convient de chercher à définir des classes plus larges de processeurs optoélectroniques d'images. Le premier perfectionnement qui peut être apporté consiste à passer du convolueur à l'automate cellulaire[15], qui a été étudié en détail ces dix dernières années sous des dénominations variées comme substitution symbolique, morphologie mathématique. [16] Le fonctionnement d'un tel automate combine une convolution à une nonlinéarité ponctuelle. Le rôle essentiel de l'optique est d'implanter la convolution, alors qu'un dispositif optoélectronique ou optique non linéaire convenable fournit la réponse non linéaire au résultat de cette dernière.

4.3. Les problèmes d'optimisation en traitement d'images

On franchit une étape supplémentaire en présentant les problèmes d'optimisation en traitement d'images dans le cadre de l'opto-informatique. Dans tout problème d'optimisation, on introduit une fonction $E(\underline{x})$ appelée énergie ou coût dont le rôle est d'évaluer l'écart entre l'image $\underline{x}$ et un idéal fixé. Le processeur doit trouver, dans tout l'ensemble des images possibles, l'image $\underline{x}_O$ qui minimise la quantité E. La définition de la fonction E prend en compte toute la connaissance disponible : image à traiter, causes de dégradation à éliminer, information apriori sur la classe d'objet concernée, et zones intéressantes. Parmi les applications classiques, on trouve la suppression de bruit, la détection de régions spécifiques, et des tâches de plus haut niveau comme la classification des motifs [17]. Des travaux algorithmiques comme ceux de Geman [18] ont montré que des fonctions d'énergie convenablement définies étaient à même de saisir, par exemple, l'information de texture, de contours ou de mouvement dans des situations réalistes et difficiles.

La charge de calcul associée, toutefois, est en général extrêmement lourde : une modification minime de l'image à traiter peut occasionner un changement important de l'énergie — on dit que le "paysage d'énergie" est erratique. Ainsi, la présence de nombreux minima secondaires empêchent les algorithmes simples de minimisation d'atteindre le minimum absolu recherché, et même de parvenir à une solution sous-optimale acceptable : pour la plupart des fonctions d'énergie utiles, le problème de recherche du minimum absolu ne peut être résolu en temps polynomial — c'est à dire en un nombre d'opération qui s'exprime comme un polynôme du nombre de pixels de l'image à traiter : l'image optimale $\underline{x}_O$ ne peut être trouvée que par une exploration exhaustive de l'espace de toutes les images possibles, dont on sait bien que le cardinal croît exponentiellement avec le nombre de pixels. En conséquence, il est en pratique impossible de trouver ce minimum absolu.

La situation est même encore plus difficile : les procédures sous-optimales efficaces connues elles-mêmes sont lourdes au point d'être impraticables. Prenons pour exemple le recuit simulé, tel que l'a proposé Geman dans le travail cité plus haut. La recherche d'une bonne solution sous-optimale requiert typiquement d'itérer une procédure de modification de l'image estimée une très grand nombre de fois, de l'ordre par exemple de quelques millions pour chaque pixel. Ceci n'est envisageable que si une implantation parallèle peut être mise au point : nous nous proposons de montrer ci-dessous que tel est précisément la tâche que l'on peut assigner à l'opto-informatique. Pour conclure cette discussion, nous nous intéressons à la mise en oeuvre d'algorithmes de traitement sur des images réalistes avec un "niveau de parallélisme optique".

5 -UN EXEMPLE SIMPLE DE RECUIT SIMULE : LE DEBRUITAGE D'IMAGES BINAIRES

Nous décrirons ici un cas simple : le débruitage d'images binaires. L'algorithme utilise un modèle markovien d'image avec interconnexion aux quatre plus proches voisins [19]. La dégradation de l'image est due à un "bruit d'inversion", qui transforme en pixels noirs certains pixels blancs et réciproquement. La fonction d'énergie convenable pour la restauration d'image est donnée par l'équation suivante :

$$E(B) = \lambda^2 \sum_i (x_i - b_i)^2 - \sum_i \sum_{j \in V_i} b_i b_j \,, \quad (1)$$

les pixels de l'image bruitée à traiter $\underline{x}$ sont notés x_i. $\underline{b}$ est l'image traitée, ses pixels sont notés b_i. V_i est l'ensemble des voisins du pixel i à prendre en compte. Tous les pixels x_i et b_i ne peuvent prendre que les valeurs +1 et -1. Cette fonction d'énergie établit un compromis entre deux termes : la différence entre l'image d'entrée et l'image traitée et l'uniformité de cette dernière. Le paramètre λ établit une pondération entre les deux termes. Pour une image d'entrée donnée, la meilleure image traitée correspond au minimum de l'énergie : le débruitage est ainsi présenté comme un problème d'optimisation.

Comme il a été expliqué au paragraphe précédent, l'exploration complète de l'espace des solutions possibles est hors de portée, et nous recommandons une méthode stochastique pour atteindre des estimations sous-optimales satisfaisantes. Dans la mise en oeuvre envisagée, cette méthode est le recuit simulé, qui a l'avantage d'être simple et d'un usage assez général.

Pour analyser cette technique, partons de l'expression de la différence d'énergie occasionnée par l'inversion d'un pixel b_i :

$$\Delta E_i = E(b_i = 1) - E(b_i = -1)$$

$$\Delta E_i = -\lambda^2 x_i - \sum_{j \in V_i} b_j \quad (2)$$

A l'aide de cette expression, on peut s'approcher du minimum recherché par une procédure itérative qui explore chaque pixel un grand nombre de fois (voir Figure 8). Par recuit simulé, chaque pixel est mis à la valeur 1 avec la probabilité suivante :

$$p(b_i = 1) = \frac{1}{1 + \exp(\Delta E_i / T)}. \quad (3)$$

Cette fonction dépend du gradient d'énergie au point considéré et d'un paramètre de contrôle appelé T, qui

reçoit le nom de température en raison de l'analogie entre l'équation 3 et les lois de probabilité de la physique statistique. Initialement fixé à une valeur très élevée, le paramètre T assure une distribution large et donc une exploration étendue de l'espace des états ; il est ensuite diminué progressivement jusqu'à atteindre zéro : dans ce cas, comme avec un algorithme de gradient, la probabilité vaut 1 si ΔE est positif et 0 s'il est négatif.

(a) (b) (c)

Figure 8. Un bruit aléatoire a été imposé à l'image initiale (a), de 64x64 pixels, en inversant la valeur de 20 % de ses pixels (b). L'image bruitée a été restaurée par l'algorithme décrit dans le texte : le résultat (c) est une bonne estimation de l'original.

6 - L'OPTOELECTRONIQUE POUR LA PARALLELISATION DU RECUIT SIMULE

Dans cet exemple, l'opto-informatique peut intervenir pour trois fonctions : la convolution pour le calcul de ΔE, la production des nombres aléatoires nécessaires pour la décision stochastique de l'équation (3), et le seuillage optoélectronique. L'exemple est en fait suffisamment simple pour que l'ensemble de ces trois fonctions constitue le processeur tout entier, qui se comprend donc les fonctions suivantes :

- la réalisation optique de la loi de probabilité par l'intermédiaire de quatre signaux optiques
- une amplification différentielle de signaux optique suivie d'un seuillage dont le résultat provoque l'allumage d'une source lumineuse pour coder l'état résultant $b_i = 1$ ou -1.

Ces opérations ont lieu en parallèle sur toute l'image, ou plus exactement sur tous les pixels dont les états n'interagissent pas par des termes communs dans ΔE. Dans le cas présent de connexions aux plus proches voisins, cela revient à faire évoluer la moitié des pixels en parallèle à chaque étape.

6.1 - La convolution optique :

Le calcul du gradient d'énergie ΔE implique souvent des convolutions, ce qui nous ramène au paragraphe 4.2. C'est le cas notamment dans notre exemple : la deuxième moitié de l'équation 2 peut se lire comme la superposition de l'image d'entrée $\underline{x}$ avec la convolution de l'image $\underline{B}$ par un noyau restreint au voisinage V^i.

6.2. Production parallèle de nombres aléatoires :

La production de grandes quantités de nombres aléatoires en parallèle avec une bonne qualité statistique est une difficulté. Il s'agit ici de fournir des nombres aléatoires indépendants correspondant à la loi de probabilité de l'équation (3) à tous les PE, ce qui représente typiquement quelque 10^{10} nombres aléatoires par seconde sur une puce microélectronique. Nous avons proposé l'utilisation à cet effet d'un phénomène aléatoire physique bien connu, la granulation cohérente (speckle) : la figure de speckle est projetée sur un réseau de photodiodes. La partie électronique de notre processeur parallèle est constituée par une puce de "pixels intelligents" comprenant au moins un photodétecteur par pixel. Avec un séquencement convenable des opérations, le même photodétecteur peut servir à la fois pour l'entrée de l'image, pour l'entrée de la valeur de ΔE qui résulte de la convolution, et pour l'entrée des nombres aléatoires.

Plus précisément, nous avons montré que les statistiques du speckle se prêtent bien à la production de la loi de probabilité voulue, et que la température T peut être simulée directement par le flux moyen dans le champ de speckle, donc par la puissance du laser correspondant [20]. Résumons le principe utilisé en quelques phrases.

Une figure de speckle "complètement développé", intégrée sur l'aire d'une photodiode, suit une loi de probabilité qui dépend du nombre de grains de speckle sur la surface du détecteur [21]. Nous avons démontré expérimentalement la production de la quantité voulue de nombres aléatoires à l'aide d'un diffuseur mobile placé devant un réseau de photodiodes sur 1 cm^2 de silicium. 22. La Figure 9 illustre l'obtention de la loi de probabilité de l'équation (3) : on recourt en fait à deux photodétecteurs. Par superposition des images concernées ou bien avec une étape intermédiaire de mémorisation, on ajoute le speckle arrivant sur un photodétecteur à la quantité ΔE. Le résultat est envoyé à l'entrée positive d'un seuillage électronique dont l'entrée négative est constituée par le speckle reçu par le second photodétecteur. Le calcul montre que par un choix convenable des paramètres dimensionnels, la probabilité pour que l'entrée positive excède l'entrée négative est une bonne approximation de l'équation (3).

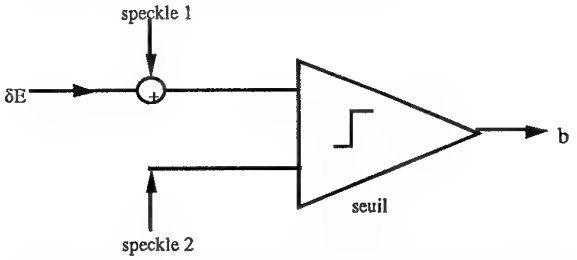


Figure 9. Génération de la loi de probabilité de l'équation (3) à l'aide de deux échantillons de speckle.

6.3. Seuillage optoélectronique :

Des matrices optoélectroniques ou optiques non linéaires de conception récentes comme les SEEDs et les photothyristors PnpN mentionnés plus haut se prêtent bien à l'opération de seuillage voulue. Nous avons publié une validation expérimentale dans le cas d'un réseau de PnpN [23]. La fonction de seuillage de la Figure 9e st réalisée dans ce cas par la paire différentielle de photothyristors optiques. Le résultat, c'est à dire la valeur du pixel b_i,, est disponible sous la forme de l'émission de la LED, ce qui se prête bien à la mise en oeuvre optoélectronique de l'itération suivante.

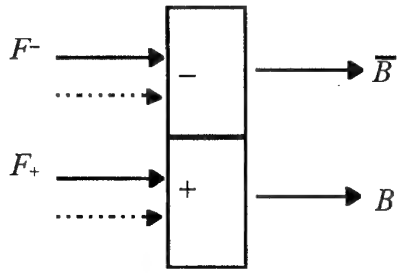


Figure 10. Utilisation d'un pixel intelligent pour le recuit simulé parallèle. Les traits pointillés représentent deux entrées indépendantes de speckle. La sortie B_i est activée avec la loi de probabilité de l'équation (3).

Illustrons les trois opérations optoélectroniques (6.1 à 6.3) dans le cas de notre exemple de débruitage. Un signal optique ayant la distribution de probabilité voulue étant créé par différence de speckles [24], un comparateur à deux photodétecteurs d'entrée et deux sorties complémentaires superposées au champ de speckle réalise l'implantation parallèle du recuit simulé. Le principe de l'opération est présenté sur la Figure 10. Au champ de speckle qui éclaire les deux entrées sont superposés les signaux notés F_+ et F_- (voir ci-dessous). Le comparateur qui suit provoque l'allumage de la diode électroluminescente du côté du détecteur qui a reçu le plus de lumière. Le pixel b_i est déterminé par la sortie active : b_i=+1 si le résultat est positif et b_i=−1 s'il est négatif.

La probabilité pour que la sortie B soit active est donnée par l'équation :

$$p(B)=\frac{1}{1+\exp\left(-\left(F_+ - F_-\right)/T\right)} \qquad (4)$$

où le paramètre T dépend de la valeur moyenne du speckle détecté.

A l'aide d'une matrice de tels comparateurs, l'algorithme voulu est réalisé si l'on peut identifier les quantité suivantes en utilisant l'analogie entre les équations 3 et 4 :

$$F_+ = \lambda^2 x_i + \sum_{j\in V_i} b_j$$
$$F_- = 0 \qquad (5)$$

Les valeurs de sortie des pixels voisins font partie des entrées nécessaires. Une connexion de la sortie d'un PE aux entrées de ses voisins est donc nécessaire. Comme les valeurs positives et négatives de b_i sont codées par des faisceaux lumineux d'intensité positive, la sommation de l'équation 5 doit être décomposée en ses parties positive et négative. Les deux entrées requises sont donc en fait :

$$F_+ = \lambda^2 x_i + \sum_{j\in V_i}\begin{bmatrix}1 & \text{si } b_j = 1\\ 0 & \text{si } b_j = -1\end{bmatrix} = \lambda^2 x_i + \sum_{j\in V_i} B_j$$
$$F_- = \sum_{j\in V_i}\begin{bmatrix}0 & \text{si } b_j = 1\\ 1 & \text{si } b_j = -1\end{bmatrix} = \sum_{j\in V_i} \overline{B}_j \qquad (6)$$

Un hologramme synthétique conçu spécialement à cet effet peut produire le motif d'interconnexion requis. Dans notre exemple, l'interaction est réduite aux quatre plus proches voisins pour limiter la complexité du motif. La Fig. 11a présente un schéma possible de mise en oeuvre. Pour plus de clarté, on n'a représenté que les connexions d'un pixel donné à trois de ses voisins mais il est entendu que le même motif d'interconnexion s'applique à tous les pixels.

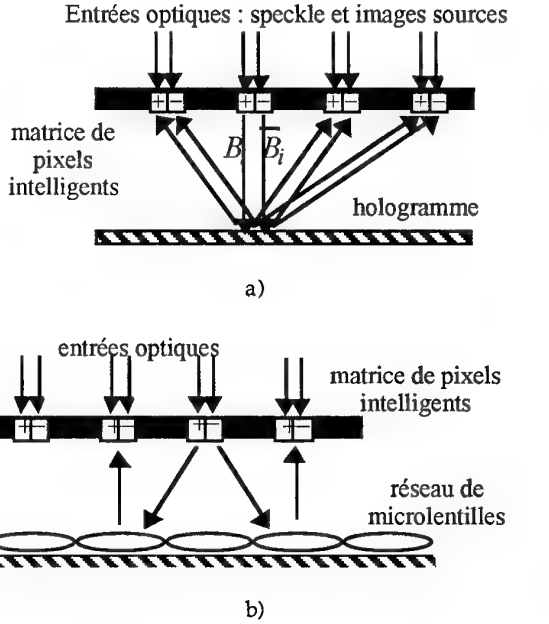


Figure 11 Implantations possibles des motifs d'interconnexion (a) par optique diffractive, (b) par optique réfractive.

Un autre schéma de mise en oeuvre possible recourt à un réseau de microlentilles destinées à envoyer la lumière

émise par un pixel vers l'entrée de ses voisins. Cette situation est schématisée sur la Fig. 11b.

On peut recourir à divers dispositifs optoélectroniques pour mettre en oeuvre cette méthode : des SEEDs, des matrices de laser VCSEL. Notre expérience concerne pour l'instant les photothyristors PnpN, qui ont l'avantage de combiner toutes les fonctions voulues de détection, amplification différentielles, seuillage et émission en un seul dispositif [25] et donc de bien se prêter à la mise en cascade des opérations. Leur fréquence de travail peut atteindre 100 kHz. Elle est limitée par la puissance optique émise et non par leur temps de réponse. Ils permette une intégration poussée, avec plus de 10^5 éléments par cm^2. Nous avons démontré leur comportement pour le seuillage de nombres aléatoires.

Une difficulté, toutefois, est que les photothyristors PnpN émettent en LED et non en diodes laser : leur lumière occupe 2π stéradians. Pour bien l'utiliser, il conviendrait de les recouvrir de microlentilles collimatrices.

7 - CONCLUSION

L'intention de ce chapitre était de souligner la mise en oeuvre de l'atout principal de l'opto-informatique pour les systèmes de traitement futurs : le nombre potentiel de canaux d'interconnexions. Il est valorisé dans les situations où le nombre d'interconnexions nécessaires est le plus large, c'est à dire les systèmes les plus parallèles. Parmi ces derniers, on a pu distinguer

- d'une part les systèmes électroniques "massivement parallèles", comprenant actuellement environ mille processeurs puissants travaillant de concert pour des tâches de calcul complexe mais apriori quelconques,
- et d'autre part les automates cellulaires optoélectroniques "massivement parallèles", avec un niveau optique de parallélisme, c'est à dire au moins des dizaines de milliers de processeurs élémentaires très frustes coopérant à des opérations répétitives de traitement d'images à cadence vidéo. Parmi les défis qui se présentent pour le développement de l'opto-informatique, on relève en première ligne l'intégration des systèmes, avec la nécessité de mettre au point des techniques peu onéreuses et fiables pour la fabrication d'éléments micro-optiques et pour leur assemblage dans des conditions de compatibilité maximale avec les systèmes électroniques d'aujourd'hui.

REFERENCES

1 - J. Jewell, AT&T Bell Labs, 1989.

2 - Optics and Photonics News, février 1993 ; E. Zeeb et al., Optical Computing, Edimbourg, août 1994, Institute of Physics Conferences Series 139, 481-485 (1995).

3 - Voir par exemple les revues Appl. Phys. Let. et J. Lightwave Technol..

4 - P. Koppa et al., J. Phys III (France), 4 2405-2411 (1994).

5 - T. Rivera et al., Appl. Phys. Let. 64 869-871 (1994).

6 - D.A.B. Miller et al., Appl. Phys. Let. 45 13-15 (1984).

7 - D.A.B. Miller, Optical Computing '94, Edimbourg, août 1994.

8 - J. Pankove et al., "a pnpn optical switch", Proc. SPIE 963, 191-197, 1988 ; K. Kasahara et al. Appl. Phys. Lett. 52 679-680 (1988).

9 - P. Heremans, communication privée

10 - voir les actes des journées d'étude sur les procédés de la micro-optique passive, Metz, avril 1995, numéro spécial de la revue Entropie à paraître.

11 - H. Dammann, K. Görtler, Opt. Commun. 3 312-316 (1971) ; J.L. Tribillon, J.E.O.S. A, Pure Appl. Opt. 3, 389-411 (1994).

12 - J.W. Parker, Optical Computing '90, Kyoto, avril 1990 ; Th. Lemoine, 10ème journée d'étude d'opto-informatique SEE/SFO, Gif/Yvette, novembre 1994.

13 - See for example Javidi B. and Réfrégier P., editors, Optical Pattern Recognition, Euro-American Workshop (Bellingham, Washington : SPIE Press), or Horner J.L., and Javidi B., editors, Special section on Pattern Recognition, Opt. Engin. 33, 1751-1862 (1994).

14 - Rajbenbach H. et al., "Compact photorefractive correlator for robotic applications," Appl. Opt. 31, 5666-5674 (1993).

15 - Taboury J. et al., "Optical cellular processor architectures," part I, Appl. Opt. 27 1643-1650 (1988), part II, Appl. Opt. 28 3138-3147 (1989). Chavel P. and Taboury J., "Binary optical cellular automata : concepts and architectures," *SPIE Proc.* **CR35** 245-265.

16 - Huang A., "Parallel algorithms for optical digital computers", IEEE 10th International Optical Computing Conference, 13-17 (1983). Huang K.S., Jenkins B.K., Sawchuk A.A., "Binary image algebra and digital optical cellular image processor design," Computer Vis. Graph. Image Proc. 45 295-345 (1989).

17 - Bouthemy P. and Lalande P., "Recovery of moving object masks in an image sequence using local spatiotemporal contextual information," Opt. Engin. 32 1205-1212 (1993).

18 - Geman S and Geman D, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", IEEE Trans Pattern Anal. Mach. Intel. PAMI6 721-741 (1984).

19 - J. Marroquin, S.Mitter, T. Poggio, "Probabilistic Solution of Ill-posed Problems in Computational Vision", J. Am. Stat. Assoc., 82, 397, 1982

20 - Lalanne Ph., "Progress towards the optoelectronic

implementation of stochastic artificial retinae," submitted for publication.

21 - Goodman J.W., "Statistical properties of laser speckle patterns," in "Laser speckle and related phenomena", J.C. Dainty, editor (Springer Verlag, Berlin), 9-74 (1975).

22 - Lalanne Ph et al., "Optoelectronic devices for Boltzmann machines and simulated annealing," Opt. Engin. 32 1904-1914 (1993).

23 - Prémont G. et al, "Optical thyristor based stochastic elementary processor," Optical Computing 94, Proceedings, Institute of Physics Conference Series 139, 67-70 (1995).

24 - P. Lalanne, G. Prémont, D. Prévost, P. Chavel, "Stochastic Optoelectronic Retinae for Vision Tasks", Optical Computing '94, Edinburgh, August 1994, Proceedings, Institute of Physics Conference Series 139, 295-313 (1995).

25 - P. Heremans, M. Kuijk, R. Vounckx, G. Borghs, "Fast and Sensitive Two-terminal Double-heterojunction Optical Thyristors", Microelectronics Eng., 19 49 (1992).

Impact of Optics on Computing Systems : from Optical Interconnects to Dedicated Optoelectronic Machines

Pierre Chavel
Institut d'Optique (Center National de la Recherche Scientifique)
BP 147, 91403 Orsay cedex, France

SUMMARY

In telecommunication systems, optics is now a pervasive technology. Its advantage of high communication throughput is relevant also to computing systems. However, eliciting the full benefit of optics will first require to develop techniques for integrating a large number of optical channel in a system and to understand the implication on computer architecture.

In this chapter, we start with a review of the physical bases that justify the use of optics for implementing interconnect networks and, more generally, for designing future computing (as well as digital signal processing) systems. This analysis determines the logical sequences of the following sections : optoelectronic technologies that are already available at present allow to demonstrate a number of functions and to extrapolate to many others. One can envisage an evolutionary path and a revolutionary path. In the first case, optical functions are added to architectures that are otherwise determined by standard microelectronics concept : under this heading, we shall examine optical interconnect networks for computing systems with a relatively high degree of parallelism. The second path implies completely revisiting architectural concepts down to circuit design. Based on the concepts of "smart pixels" and cellular automata, it suggests a number of ideas for dedicated processors applicable in particular to vision machines with massive parallelism — we shall comment in due time on the particular meaning of the latter expression in the context of optical and optoelectronic computing.

1 - WHY OPTICS ?

It is appropriate to remember at this point that the functions necessary for the operation of any digital computer can be reduced to only three primitive operations : binary logic, point to point transfer of information — i.e., interconnection —, and memory. In this chapter, we shall consider only the first two, while the contribution by S. Esener will cover the last one as well. It is straightforward to derive from physical principles the arguments in favor of using optics, and more precisely of combining optics with electronics for the benefit of computing system performance : these rely on speed — or more accurately on time delay required to perform one primitive operation —, on the communication bandwidth (in the time domain), and on the density of interconnect (in the space domain).

1.1 - The "speed" of optical functions

It is commonplace to state that the interest of using light in a computer stems from the processing speed. This statement, however, deserves some comment so as to avoid any naive oversimplification. Specifically, it is not automatic that either a logical or an interconnect operation is performed faster by optical means than by electronic means.

Optical logic essentially relies on the effect of electromagnetic fields on the energy bands shapes and populations in solid state devices. These phenomena are exactly the same as those used by electronic logic. The particularities of some materials or some component family may determine a certain advantage of some device, but only relatively small factors come into play here so that a complete change of technology is not warranted in favor of optoelectronic solutions. For example, the record switching time for a transistor is of the order of a picosecond, just like that for an optical bistable element.

However, it is known that computers are presently limited by interconnect delays rather than by switching times. It is true that the speed of electrons in a conductor does not exceed the order of km/s, while the speed of light in a vacuum is close to 3.10^8 m/s (or, in suitable units for the scale of a computing system, 30 cm/ns). It must be reminded here that his comparison, however, is completely irrelevant.

The speed of signal is not in itself an advantage for interconnects. To send an information by electrical means from A to B does not imply to physically move a charge carrier from A to B because the message is not carried by the electrons (or holes). Instead, the message is contained in the electromagnetic field produced by the carriers. That field is of the same nature and propagates at the same speed as the optical field : only the frequency is different, the optical frequency range being around hundreds of terahertz ($1\ \mathrm{THz} = 10^{12}$ Hz). On closer look, it is true that the relevant speed is not that of electromagnetic waves in a vacuum, but the group velocity in the specific medium, which depends on its refractive index and dispersion : in usual materials, there may be a slight advantage in favor of optics. But this is just a small bonus, not a decisive argument.

Paper presented at the AGARD SPP Lecture Series on "Optical Processing and Computing" held in Paris, France from 12-13 October 1995; Rome, Italy from 16-17 October 1995 and Ankara, Turkey from 19-20 October 1995 and published in LS-199.

1.2 - Propagation delay and communication delay

There is still more to that discussion, however, because the propagation time of an electromagnetic wave is not identical to the time it takes to transmit an information unless suitable detection means are available to perceive the signal as soon as it arrives, which implies extremely high sensitivity. Individual photon counters at optical frequencies and beyond do exist, and extremely sensitive electromagnetic detectors are being developed for all frequency ranges. However, they are cumbersome and expensive. The relevant question in computer design is not to know whether a signal has been transmitted, but whether the amount of energy transmitted to the destination is sufficient to reach the detection threshold of its detector. At the scale of a computing system in a rack as opposed to long distance communications, the propagation time is small compared to the time needed to send that amount of energy. In practice, transmitting energy on a conducting line implies to bring an electrode to a given reference potential through a given impedance. That impedance is mainly due to the capacitance between the various lines in the circuit. Save for the immediate neighborhood of the light source and detector, optical beams to not necessitate any conductor and this is therefore a substantial gain in favor of optical communication. Whether this is practically beneficial in a given situation implies a careful balance of the energy associated with emitting and detecting light, which in turns depends on source and detector technology and on integration techniques : these issues will be considered below.

1.3 - Information channel capacity

Whenever an information is carried by an optical beam, a light source is modulated by the signal to be transmitted. This modulation is superimposed on the optical carrier frequency, which as already mentioned is quite high. Optical telecommunications make an increasingly good use of the huge bandwidth available around the optical carrier, and it is justified to transpose that idea to the short distance communications that arise inside computers. Here also, only the cost and volume of practical solutions for multiplexing, detecting and demultiplexing information in a broad bandwidth are relevant to make the decision of using optics in a given context.

1.4 - Interconnect density

A final important and clear argument in favor of optical communications is the density of the interconnect network, which is to space as bandwidth is to time. The trivial argument that "photons" cross each other without interaction, so that optics can produce images with million of independent pixels is fundamentally correct. Physics allows to envisage such a considerable density of optical interconnects that the limit is out of reach in the present technological context. The implementation of optical interconnect networks through free space is therefore a major goal for optical computing. It is presently impeded by system design issues rather than by fundamental limits : it is therefore appropriate to be imaginative with cheap, compact and efficient optical interconnect designs.

1.5 - Recapitulation

To conclude this discussion, the decrease in interconnect capacitance, the bandwidth available around the optical carrier frequency and the potential density of optical interconnection networks are the clear physical benefits that advocate for optoelectronic computing. Optical logic, as well as the possibility to reconfigure networks may play a role in some cases — this appears to be true in particular in the context of all-optical switching in telecommunications. At the level of computer systems nevertheless, these issues are not of primary importance. The progress of optoelectronic computing is presently determined by device and integration technology, as will be examined in the following section.

2 - OPTICAL COMPUTING TECHNOLOGIES

2.1 - Active devices

2.1.1 - Semiconductors for optical computing

While several technological families may deserve consideration, we shall concentrate mainly if not exclusively on compound semiconductor devices, because of their remarkable progress in the recent years and because they offer a particularly good compromise between optoelectronic and conventional electronic properties. This choice is in part arbitrary and is dictated by the necessity to limit the scope of the chapter. The basic advantage of these materials comes from the combination of two factors : a favorable energy band structure and the possibility of applying to them the results of many years of technological experience developed for silicon circuits.

The dominant position of semiconductors, and in particular of silicon, in computer circuits is well known. There is no clear optical counterpart to the role played by the transistor effect in those circuits : in optical computing, one sensible option is to use transistor for all logical operations and to hand over to optics for data transmission. Then, adequate solutions are needed for the emission, modulation and detection of light beams. In spite of some progress in recent research and although light detection is easy on silicon, the indirect bandgap structure of this material is a handicap : there is presently no efficient and well developed solution for the

emission or modulation of light in silicon technology. Compound semiconductors are therefore a better solution. These belong in particular to the III-V and II-VI families, whose names derive from the position of their component elements in the periodic table. One important case is gallium arsenide, a direct gap semiconductor of the III-V family that has interesting properties for the emission and modulation of light and whose technology has been well investigated for reasons independent of optics. In association with silicon or in independent monolithic circuits, GaAs is therefore the main material for optical computing. It can be used in bulk form or, with suitable high technology tools, in the form of alloy layer stacks with variable composition where other elements of the III and V columns are used together with gallium and arsine : the design of such "heterostructures" gives a considerable flexibility to shape spectral properties at will and to optimize performances.

The physics of such components relies in all cases on the excitation of carriers from the valence band to the conduction band and on the effect of a static field or an incoming illumination on the band structure. These effects modify the absorption and emission spectrum. They will not be described here, but the specifications of some current devices suitable for optical computing applications will be given.

2.1.2 - Semiconductor lasers

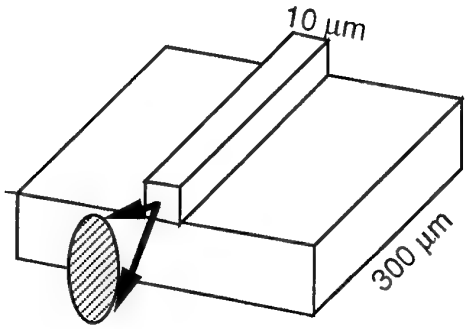


Figure 1. "Horizontal" cavity laser.

One of the most important and common devices is clearly the laser diode, which in fact was developed for a completely different domain of application. The most usual structure is schematically depicted on Figure 1 : the diode length is relatively large, on the order of 100 µm, so that the amplification is sufficient to reach the lasing threshold. Recent progress in the design and growth of semiconductor layer stacks have made it possible to reach the lasing threshold with significantly smaller cavities and thereby elicit the "vertical" emission of light — the word "vertical" is used here to designate the direction perpendicular to the substrate — so that two-dimensional integration becomes convenient. Figure 2 sketches a typical array of vertical cavity surface emitting lasers (VCSELs).

When the first VCSEL array was announced by Jewell [1], great expectations arose in optical computing. For the sake of demonstration, the first chip contained 2 million lasers per square centimeter, i.e. the same integration density as for transistors on a memory chip. However, it is worthwhile to note that the individual addressing using the same number of pairs of electrodes was, and still is, out of reach of technology : a common ground plane was provided inside the structure, but to switch one laser on an individual contact by a test pin was required to provide the supply voltage : only optical interconnects can be expected to perhaps allow the individual addressing of so many channels in such a small area. Since that first demonstration, VCSEL arrays have been developed with individual addressing : they consist of a few tens of elements and are becoming commercially available. Some performances are mentioned in reference 2.

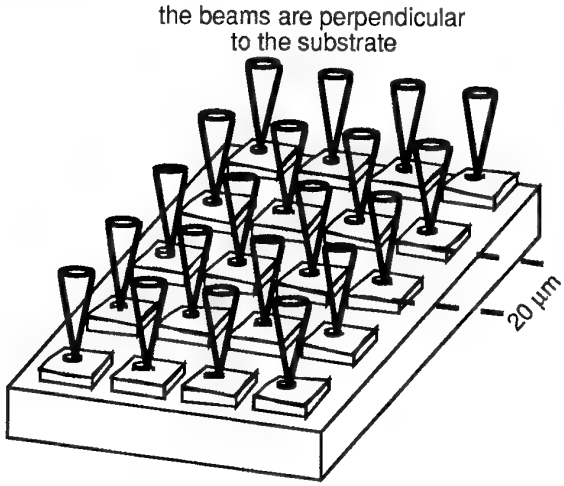


Figure 2. Vertical cavity laser array.

2.1.3 - Modulators

Laser beams may be modulated either internally, through the excitation current, or externally, using an external device. The latter solution is particularly well adapted to very high bandwidth, with the present records in the tens of gigahertz [3] for telecommunication applications. The gigahertz range has already been reached in commercial long distance applications. Whether the same can be practically applied to high rate interconnects in computer systems will depend on the complexity and cost of the circuitry required to multiplex and demultiplex many smaller bandwidth signals.

The issue of the choice between light emission on the chip or modulation of an external beam is still open. Modulation, indeed, requires light to be emitted out of

the chip by a laser that plays the role of a power supply for optical energy. It has however the advantage of a smaller consumption on the modulating chip. Roughly speaking, the modulator has the same structure as a laser diode, but it is used in inverse polarization and has therefore a high impedance. An example of performances can be found together with an example of application in reference 4.

2.1.4 - Bistable elements

A higher functionality than modulation can be found with logical optical elements. Often misleadingly designated as "optical transistors", they share with the transistor the property of having two well defined and well separated states ; the reverse polarity conductivity phenomenon known as the transistor effect, however, does not come into play. A command beam of power P is absorbed and modulates the transmission T or reflectivity R of the device, which can therefore be considered as a light switch, an optically driven shutter. The most common case is the optical bistable element, whose R(P) characteristic curve shows a loop similar to the hysteresis cycle of magnetic materials. Reference 5 describes the present record bistability threshold under continuous illumination.

2.1.5 - Integrated logical optoelectronic circuits

We shall mention in particular two components that can be used to make circuits : the SEED device and the PnpN optical photothyristor.

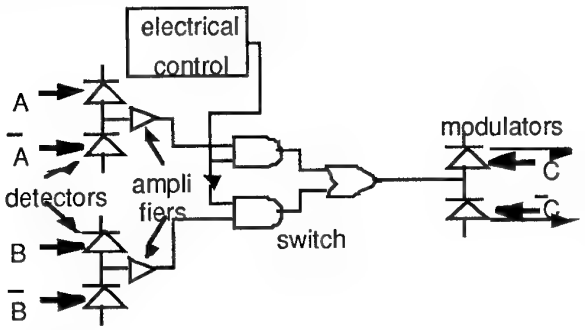


Figure 3. One FET-SEED based optoelectronic circuit that can be integrated ; the circuit contains three pairs of SEEDs arranged into pairs, two of which are used as optical data inputs and one as optical data output.

The self-electro-optic effect device (SEED) was first introduced by Miller of ATT Bell Laboratories in 1985 [6]. Its structure is designed in such a way that an applied voltage lowers the gap through an electro-optic effete, so that under suitable illuminating wavelength the device switches from transparent to opaque (with a given switching contrast). SEEDs can be associated to a field effect transistor that amplifies its photocurrent : as shown on Figure 3, SEEDs can then be part of fairly complex logic cells. Some current specifications are a switching energy of 20 to 30 fJ and a maximal seed of up to 650 Gbits/s (which is more than any realistic circuit can handle) [7]. ATT SEEDs are used in external collaboration for making demonstrators (but at this time no commercially competitive systems yet). The association of microelectronic functions on gallium arsenide together with bistable modulators of the SEED family is one major example of the present state of development of "smart pixels" : smart pixels are logical cells with a "smartness" derived from the presence of one or more optical input or output gates on the circuit surface — alternatively, they can be considered as optoelectronic modulators and detectors with a "smartness" derived from the presence of an associated logic circuit.

In 1988, two teams have simultaneously proposed a light emitting thyristor structure [8]. One result is the development by IMEC in Belgium of PnpN optical photothyristors, that have an optical input, an electrical input and an optical output. Like in the SEED, the optical input is the photosensitivity of the device, in this case in the gate area, the electrical input is the voltage applied to the gate, but here the optical output is in the form of an light emitting diode (LED) constituted by the thyristor itself in its "on" state. When two such thyristors are paired together in parallel and wired in with a common load resistor in series, they behave as a differential thresholding amplifier : applying the driving voltage will cause only the element in the pair that received more light to switch on. A record differential photodetection energy threshold has been obtained [9] : 20 fJ/μm^2 (the required electric energy is not included in the figure).

2.2 - Micro-optics

2.2.1 - The optoelectronic integration challenge

With devices such as those mentioned in section 2.1, it can be said that the set of active devices for optoelectronic functions is now reasonably rich, at least at the laboratory level. Their implementation in computing systems now requires two complementary research efforts : system case studies to assess their practical interest and system integration studies for a reliable and realistic implementation. Before we cite some demonstration systems, we shall first discuss the latter aspect.

One frequent and serious argument against optical computing is the need to align components with an accuracy of the order of the device size. To increase density and to reduce power consumption automatically means to decrease size. Optical interconnections between two devices obviously request careful alignment : if for example a SEED device has 10 µm square input windows, the illuminating beam must obviously be

aligned to better than 10 µm accuracy. The solution to the alignment problem adopted in electronic systems is to introduce a hierarchy of interconnect levels : on chip, the accuracy is guaranteed by lithography and the devices can be quite small. In hybrid circuits, the bonding is made during fabrication. In electronic backplanes, boards have usually 2.54 mm spaced interconnects that are defined with a low accuracy and can therefore tolerate variations in the positioning of the board of several tens of a millimeter : the associated significant loss in compactness is the price for practical assembly techniques. Similar practical constraints apply to optical interconnects. To benefit from their potential density, it is therefore necessary to develop assembly, alignment and integration techniques applicable to light beam manipulation : this explains the need for micro-optics applied to optical computing.

There are three main types of beam manipulation functions :
• deflection, which is typically performed using mirrors or prisms ;
• focusing, which is made by lenses or spherical mirrors ;
• beam splitting, as can be done by partially reflective mirrors.
These functions, or course, are implemented through the use of the three standard functions of passive optical components : refraction, reflection, diffraction.

2.2.2 - Micro-optical components :

Refraction and reflection, as described by the well known Snell's laws, allow to control the direction and the focusing of a beam through a proper distribution of refractive index, typically using plane or spherical dioptres. The fabrication of lenses, mirrors and prisms is usually made piece by piece (perhaps in batches) and requires cutting and polishing steps. The fabrication of two dimensional arrays of such elements is not compatible with individual polishing of elements and therefore requires to devise new processes such as ion exchange in glasses or reactive diffusion of vapors in polymers [10]. Accurate control of shape by these methods is a new challenge for the opticist : it is required to guarantee beam quality and therefore focusing on a integrated set of devices.

It is known that diffraction sets a fundamental limit to the size of a light spot : if a beam of wavelength λ has a cross section D in some plane (P), its direction cannot be defined to better than λ/D. It follows that if it is focused by a lens at a distance f from plane (P), its diameter cannot be smaller than $\lambda f/D$. More accurately, the minimum spot size is a function $V(\lambda,D,f)$ that asymptotically reaches $\lambda f/D$ if D is much larger than f and asymptotically reaches λ if D is much smaller than f. Diffraction, however, is not just a limitation. It can be used to pattern a surface into elements of a small size D is such a way that the combination of the diffracted beam will generate by interferences a given result. This is a new degree of freedom that can be used for the shaping of light wavefronts. For example, one can fabricate lenses by planar techniques, or divide an incoming beam into three or more beams with a prescribed distribution of intensity. Figure 4 shows the uniform illumination of 16 spots through a diffraction grating with a suitable groove profile. Such gratings are known as Dammann gratings [11]. This is the domain of diffractive optics, which is also known under two other names : because of a striking analogy with the principle of holography, diffractive optical elements are sometimes considered as one case of computer generated holograms. In addition, the fabrication of diffractive optical components makes an increasing use of lithography techniques that were developed for microelectronics : the fabrication of some elements in successive layers has suggested the name "binary optics", which is somewhat misleading in this context.

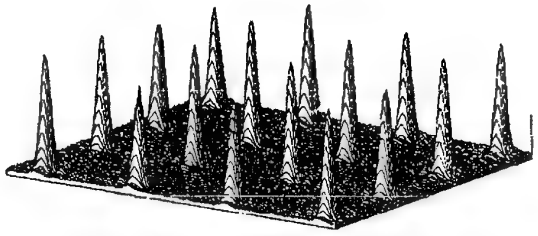


Figure 4 Illumination distribution generated by a 16 points array illuminating grating.

3 - SOME EXAMPLES OF OPTICAL INTERCONNECTS

In this section, we describe three recent cases. The first is characteristic of the work developed by Thomson CSF in ESPRIT project Olives and Holics and is close to industrial needs. The other two are further ahead and result from collaboration between our laboratory and various partners.

3.1 - Optical backplane

The idea with this project [12] is to increase the number of interconnect channels between boards in one electronic cabinet with N boards plugged in a backplane connector (N is typically of the order of 10). The various board are connected to each other through a circuit placed perpendicular to the boards (see Figure 5). Each connector may typically include 200 electric pins and is limited by crosstalk modulation at about 100 MHz. Optics provides additional links through one additional connection unit that consists of one optical input and one optical output. An optical circuit (OC) is placed

perpendicular to the boards like any backplane connector : the usual geometry is therefore preserved and augmented by optical possibilities. The optical output at each board consists of one laser, while the optical input consists in a set of N-1 photodetectors, all aligned at the border of the board. To each laser and to each photodetector on the boards corresponds one hologram on OC. Light from the laser on board i (i = 1 through N) reaches immediately its associated holographic coupler HC_i on OC. Circuit OC is made on a glass plate and light is sent by the hologram inside the substrate, where it is totally internally reflected as in an optic fiber, until it reaches an holographic outcoupling element HD_{ij} located next to one of the photodetectors of board j=i+1 or j=i-1. Part of the light is extracted by HD_{ij} and sent to its detector, while the rest of the light propagates on to the next board. With N coupling holograms and N(N-1) outcoupling holograms, a complete interconnection between the N boards is obtained. The communication throughput is limited only by the on-board modulation possibilities, i.e. by the complexity of the electronic steering unit.

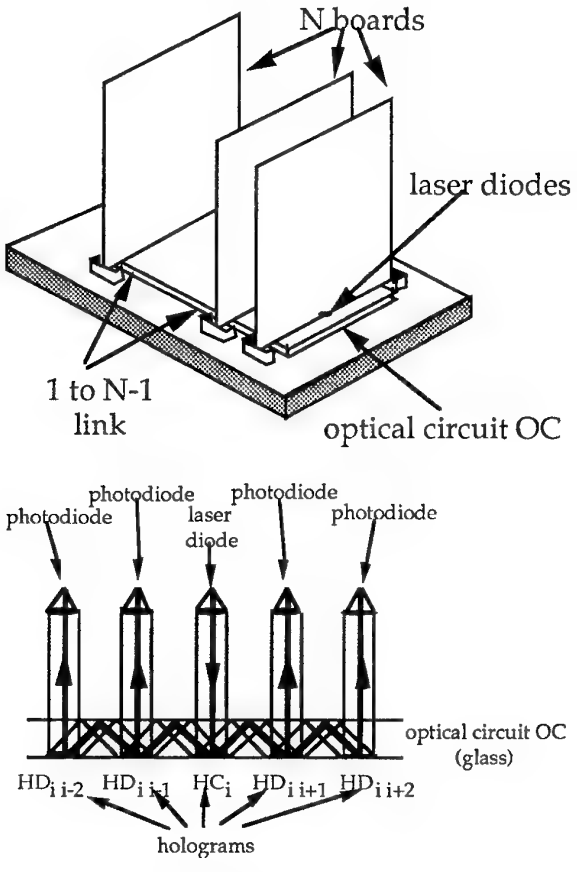


Figure 5 Optical backplane connector for an electronic cabinet (top : perspective ; bottom : cross-section).

3.2 - Address header recognition for all optical communication

This experiment is the final demonstrator of the project on Optoelectronic Matrices for Signal Processing funded by the French Ministry of Research (1987-1994). The goal is to illustrate the possibility of parallel processing with novel nonlinear optical and optoelectronic devices developed by the partners in the project. The demonstrator is a 6 bit address decoder. An incoming optical signal carries a 6 bits destination header (see Figure 6). The optical identification must allow to direct the beam into the proper output channel out of $2^6 = 64$ channels. The implementation relies on two main active components : one 8x8 array of multiple quantum well electro-optic modulators made by Thomson CSF and an optical bistable plane made by CNET, France Telecom, Bagneux, both in gallium arsenide technology. The demonstration is intended to lead to further development in the domain of telecommunications as well as of computer systems.

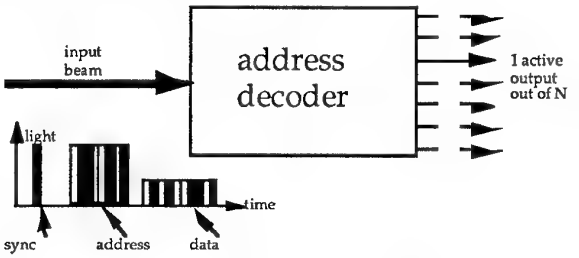


Figure 6 Packet header decoding.

The principle of operation is as follows (see Figure 7). The bistable elements are initially set to their high reflectivity state through 64 holding beams generated by a Dammann grating array illuminator (see above). The signal beam, itself divided into 64 equal parts, illuminates each pixel of the electro-optic modulator array. Each modulator receives the same sequence : a synchronizing pulse is followed by the binary destination address, encoded as complemented data (bit "0" is encoded as a bright-dark sequence, and bit "1" by a dark-bright sequence), and then by the data to be transmitted to that address. Upon reception of the synchronizing pulse, each modulator emits a sequence of 12 binary electronic pulses that represent their own address in an inverted complemented format. It results that of the 64 beams, only one never exceeds the bistability threshold of the bistable elements : this is the destination channel. The other 63 bistable channels are switched down to their state of low reflectivity. The data in the message are encoded with a lower intensity than the address so that the bistability threshold will not be exceeded on the destination channel.

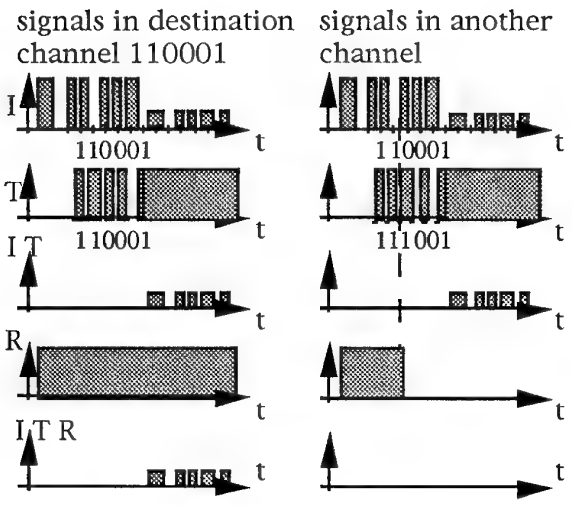


Figure 7 . Principle of the address recognition logic :
I = input light. T = electro-optic modulator transmission.
R = bistable element reflectivity.

In the present status of the experiment, the header address is recognized at a clock speed of 10 MHz (600 ns are therefore required for the full address), but it is limited by electrical equipment constraints rather than by the possibilities of the optoelectronic or non linear optical devices. However, surface inhomogeneities issues and imperfect wavelength adaptation of the elements limits crosstalk to an unsatisfactory level.

3.3 - Optical switching through active beam steering

In a recent joint study with INRIA, (Rocquencourt, France), University of Jerusalem and Supelec (Metz, France), we have experimentally illustrated an active beam steering operation. The final goal is to implement an optical "switch cube" that will allow to set an arbitrary interconnection pattern among N cabinets, with N of the order of a few tens.

The present performances were dictated by the amount of funding available and are limited to the possibility of sending a 100 Mbits/s signal from site A either to site B or to site C or to both B and C. The technology selected, determined by commercial availability, is acousto-optic deflection. An acousto-optic crystal is a device capable of steering a light beam through an acoustic driving signal, that itself arises from an electric signal through a piezoelectric transducer. One can either direct the incoming beam globally to one direction or separate it into several beams with arbitrary direction within a range of a few degrees.

The long term perspective is much more ambitious. An optical switch cube will consist of N data input ports, N address input ports and N data output ports and will be able to provide an arbitrary interconnection between N sites. Each data input and each data output itself consists of M optical fibers that typically carry the M bits of one word in some suitable format. Each address input is a N bit code arising from one of the N sites, the "1" bits designate the addressees of some message. When site i (i =1 through N) needs to send a message, it first sends an electrical N bit code to the i-th address port of the switch cube so as to identify the addressees. After a fixed delay, typically 1 microsecond, determined by the acoustic propagation speed in the crystal, the proper connection is established and site i then sends the data : M lasers are modulated at the appropriate clock frequency in a bit parallel mode and illuminate the M input fibers of the i-th input port. At the output, after a latency that is strictly limited to the propagation distance, the addressees receive the signal through M output fibers.

4 - CLASSES OF PARALLEL OPTOELECTRONIC PROCESSORS :

4.1 - Optical scale parallelism :

Aside interconnects, the processing of real images is one area where optical computing may come up with competitive solutions. The basic reason, as with interconnects for electronic computing systems, is the number of connections that are needed before any sensible processing task is completed ; it is particularly large in parallel processors working on images, which is where free space optics is attractive.

The word interconnect here should be understood in a broader sense than above, including data input and output as well as the provision of various control signals to the processing elements and of course exchange of information among the processors themselves. We are interested in "optical scale parallelism", where the processor consists in a number of processing elements (PE's) equal to the number of pixels in the image, typically at least 10^4 : this situation puts the heaviest weight on interconnects but alleviates considerably program and data transfer control. For easy image format input, all PE's should fit together on a chip a few square centimeters on a side. But technology will allow to integrate only fairly weak PE's on such a small area. We are then faced with a double challenge :

- devise optoelectronic architectures that make the best use of optical interconnects to maximize the computing power
- and find algorithms that can use it for meaningful image processing tasks.

This kind of approach, admittedly, does not directly fit in the context of architecture and design of parallel computing systems. Instead, it is aimed at the possible emergence of a different set of computing systems, that would be highly specialized in image processing tasks but compact and powerful in that context. We introduce our approach through comparison with the well known

architectures of optical correlators (i.e. convolvers) and optical cellular automata.

4.2 From optical convolutions to optical cellular automata

The simplest and most famous application of the above concept is the optical convolution, where the electronic part of the PEs reduces to photodetectors and where weighted optical interconnects that define the impulse response do all the processing. The main application is pattern recognition. The double challenge that we just mentioned then takes on the following form : can convolution be helpful in real pattern recognition problems and if yes, can optics implement such convolutions ? Progress on filter reprogrammability, adaptivity to the input signal and invariances has been fast in the last few years [13]. Also, nice implementations of rugged and compact optical convolvers have been published [14].

Convolution nevertheless shows only limited generality and it is important to seek broader classes of possible optoelectronic image processors. The next simple case is cellular automata [15], that have been investigated in some detail for about ten years under various names, including symbolic substitution and mathematical morphology [16]. Their operation cycle consists in the combination of one convolution and one point nonlinearity. The main role of optics here is to implement the convolution part, while optoelectronic or nonlinear optical devices located at every pixel respond nonlinearly to its result.

4.3. Optimization problems in image processing

One further step, then, is to introduce optimization problems on images into the realm of optical computing. In an optimization problem, an energy function $E(\underline{x})$ is introduced as a measure of the departure of an image $\underline{x}$ from an ideal goal. The role of the processor is to find the particular image $\underline{x}_O$ that minimizes function E. The definition of function E incorporates all relevant knowledge, i.e. the input data but also, for example, the sources of degradation to be removed, a priori information on the class of object, and the features of interest. Typical applications include noise removal, feature detection, as well as higher level tasks such as pattern classification [17]. Previous algorithmic work, notably by Geman et al. [18], has demonstrated energy functions that can detect, for example, texture, edge or motion in fairly realistic and difficult situations.

However, the computational load is usually extremely heavy because a small change in the image can generate a large change in the energy - the "energy landscape" is said to be wild - so that secondary minima will prevent deterministic descent algorithms from reaching the desired minimum or even an acceptable suboptimal solution. With most energy functions, the problem is non-polynomial in time, i.e. the optimal image $\underline{x}_O$ can be found only through exhaustive search in the space of all possible images, whose size increases exponentially with the number of pixels. As a consequence, it is impossible in practice to find the absolute minimum.

But the situation is even worse than that : efficient suboptimal procedures are themselves hardly practicable. Let us take the example of simulated annealing, that was advocated by Geman in the work cited above and will be developed below. Finding a good suboptimal solution will typically require to loop through a procedure of energy updating a quite large number of times, typically of the order of a few million times the number of pixels. This is still impractical unless some way can be found to implement them in parallel : as we shall illustrate now, their parallel implementation is where, in our opinion, optics may have a new role to play. In conclusion to this discussion, among the powerful algorithms that have been found to progress in the processing of real images, one subset may be open to "optical scale parallelism" and this is what we are investigating.

5 - ONE EXAMPLE OF SIMULATED ANNEALING : NOISE CLEANING IN A BINARY IMAGE

As a simple application, we selected the problem of noise removal in binary images. The algorithm involves a Markovian image model with a four nearest neighbors interconnection pattern [19]. The degradation of the image is modeled as "inversion noise" — i.e. white pixels that turn dark and dark pixels that turn white, the energy function describing this restoration process is

$$E(B) = \lambda^2 \sum_i (x_i - b_i)^2 - \sum_i \sum_{j \in V_i} b_i \, b_j \,, \quad (1)$$

where X is the noisy input image composed of pixels x_i, B is the restored output image; its pixels are denoted b_i. V_i is the set of neighbors of pixel i. All pixels can on take values of either +1 or −1. This energy function balances two terms : the difference between the input and the output and the uniformity of the restored image. The parameter λ weighs the relative importance of the two terms. For a given input image, the best restored image corresponds to the minimum of the energy function. Noise removal can thus be achieved through an optimization problem.

Since, as explained above, the complete exploration of the solution space is impossible, we use a stochastic technique to produce accurate estimates of the optimal

solution. In this implementation, we use simulated annealing for its simplicity of operation and its applicability to a vast class of problems.

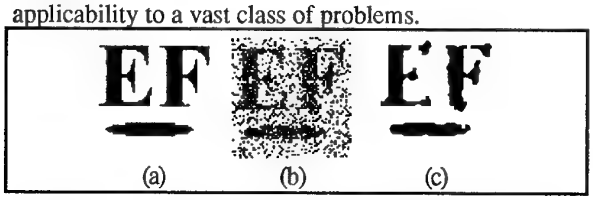


Figure 8. - A random noise is added to the 64x64 original image (a) by changing the value of 20% of the pixels (b). This noisy image is restored with the algorithm described. The result (c) is a good approximation of the original image.

To implement this optimization technique, the difference in energy between the two possible values of pixel b_i is needed:

$$\Delta E_i = E(b_i = 1) - E(b_i = -1)$$

$$\Delta E_i = -\lambda^2 x_i - \sum_{j \in V_i} b_j \qquad (2)$$

This expression can be used to minimize the energy through an iterative process that explores each pixel a large number of times. An example of this process is shown on Figure 8 In the simulated annealing algorithm, a pixels are set to 1 with the following sigmoid probability :

$$p(b_i = 1) = \frac{1}{1 + \exp(\Delta E_i / T)} \qquad (3)$$

This probability function varies with the energy gradient associated with the target pixel and with the control parameter T, which by analogy with similar probability laws in statistical physics is called the temperature. This parameter is initialized at a high value to produce a wide distribution and it is then slowly decreased until it reaches zero where the probability will be 1 for positive ΔE and 0 for negative ΔE.

6 - OPTOELECTRONICS FOR PARALLEL SIMULATED ANNEALING

In this context, optical computing can provide three functions : convolution for the calculation of δE, production of the random numbers requested to implement the stochastic decision part of simulated annealing, and optoelectronic thresholding. In the example described in section 5, these three functions together constitute the whole processor, which consists of

- the optical production of the required probability distribution, that is in fact created as a difference of two signals
- a differential pair of optical detectors that will function as a threshold to switch on a light source coding for the resulting state $b_i = 1$ or -1.

Of course, these operations are performed in parallel over the entire image — or, more precisely, over all pixels that do not affect each other's ΔE. With nearest neighbor interconnects, this amounts to half of the pixels being able to operate in parallel.

6.1 - Optical convolution :

Firstly, the calculation of energy variations ΔE often implies convolutions, and we are back to section 4.2. This is in particular the case in our example where it is clear that the lower half of Eqn 2 consists of one direct input image x superimposed with one convolution of image B with a kernel that extends over the appropriate neighborhood.

6.2. Parallel generation of random numbers :

The parallel generation of a large amount of random numbers of a good statistical quality is a problem. The requirement here is to provide independent random numbers with the appropriate statistics to all processing element, i.e. every pixel, at every energy update operation, which means around 10^{10} random numbers per second on a microelectronic chip. We have suggested to use a physical random number generator for this purpose and investigated the use of laser speckle projected onto an array of photodiodes. The electronic part of our parallel processor therefore consists of a "smart pixels" chip with at least one photodetector per image pixel being processed. Of course, with suitable sequencing, the same photodetector may be used for image input, for the input of ΔE, and for the speckle input.

Specifically, we have shown that speckle statistics can be easily molded into exactly the required form of probability law, and that the simulated temperature can be controlled directly by the average speckle brightness, i.e. the laser power [20]. Let us just summarize the basic principle in a few lines.

A fully developed speckle integrated over the area of a photodiode obeys known statistics that depend on the number of speckle grain over the detector area [21]. We have experimentally demonstrated the possibility of producing 10^{10} random numbers per second by projecting speckle from a suitably moving diffuser onto a 1 cm^2 silicon photodetector array 22. Figure 9 illustrates how to obtain the probability law required by Eqn (3) : two speckle photodetectors are used instead of one. An analogue adder combines the signal from the first photodetector with the energy difference ΔE. The result is sent to the positive input of a thresholding gate, while its negative input receives the second speckle photodetector signal. Analysis shows that, with suitable speckle parameters, the resulting probability that the

positive input exceeds the negative input quite well approximates Eqn (3). Temperature is emulated by the average speckle intensity.

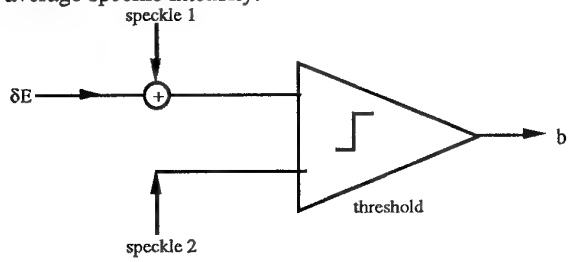


Figure 9. Generating the probability law of equation 2 with two speckle samples.

6.3. Optoelectronic thresholding :

Finally, novel optoelectronic or nonlinear optical arrays such as SEEDs or PnpN photothyristors may be used to make the required decision. Experimental validations in the case of a PnpN array have been published [23] the role of the thresholding gate of Figure 9 can be played by a differential pair of optical photothyristors. The output, i.e. estimated pixel b_i, is then available in the form of an optical signal for some further processing step or for the output of results.

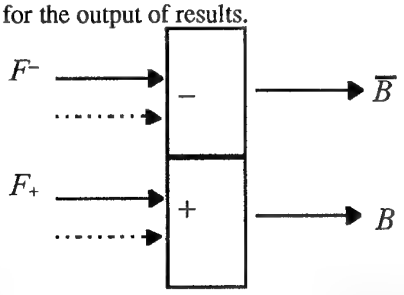


Figure 10. Use of the smart pixel for simulated annealing. The dotte lines represent two independent speckle inputs. The output B_i is active with the sigmoid probability of Eq. 3.

Let us now illustrate practically these three points (6.1 through 6.3) on the case of our example. An optical signal having the required probability distribution can be created by using a differential detection of speckle patterns [24]. A comparator with two input windows and two complementary outputs together with a homogeneous speckle field can generate the statistics required for the parallel implementation of simulated annealing. The principle of operation is simple. A speckle field (shown with dotted arrows on Figure 10) is incident onto the two inputs gates of the element, and two additional signals, F_+ and F_-, to be described below, are incident onto the appropriate inputs. The element needed is a comparator, so the output corresponding to the most intensely illuminated input is activated and emits light. The output value of the pixel b_i is defined by whichever output is active: b_i=+1 for the positive output and b_i=−1 for the negative output.

The probability to have the output labeled B active is described by

$$p(B)=\frac{1}{1+\exp\left(-\left(F_+ - F_-\right)/T\right)} \qquad (4)$$

where the parameter T depends on the detected speckles mean power.

Using an array of comparators, we use this stochastic algorithm to implement binary image restoration. It follows from the similitude between Eqs 3 and 4 that the probability function needed to solve the restoration problem with simulated annealing might be generated by using the following inputs :

$$F_+ = \lambda^2 x_i + \sum_{j\in V_i} b_j \qquad (5)$$

$$F_- = 0$$

The output values of the neighborhood pixels are among the inputs required. Some kind of interconnection must then link the outputs of each pixel to its neighbors' inputs. However, as positive and negative values of b_i are coded with the light beams positive intensity, the summation of Eq. 5 must be separated into its positive and negative parts. The two inputs needed are finally :

$$F_+ = \lambda^2 x_i + \sum_{j\in V_i}\begin{bmatrix}1 & \text{if } b_j = 1\\ 0 & \text{if } b_j = -1\end{bmatrix} = \lambda^2 x_i + \sum_{j\in V_i} B_j$$

$$F_- = \sum_{j\in V_i}\begin{bmatrix}0 & \text{if } b_j = 1\\ 1 & \text{if } b_j = -1\end{bmatrix} = \sum_{j\in V_i} \bar{B}_j \qquad (6)$$

A specifically designed computer generated hologram (CGH) can produce the required interconnection pattern. In our application, the interaction between neighborhood pixels is limited to the four nearest pixels in order to reduce the interconnection complexity. An example of this implementation is shown on Fig. 11a. For clarity, only the interconnections between one pixel and three of its neighbors are shown but it is implied that all pixels are copied onto their respective neighbors.

An alternate way to produce the interconnection pattern is to use an array of microlenses to couple the light emitted from the neighbors output into the input of the target pixel. A schematic example is shown on Fig. 11b. With the illumination of the entire matrix by a speckle field and when the input image X is entered in the system, the probability criterion of Eq. 3 determines the output of every pixel, with the energy gradient expression of Eq. 2 .

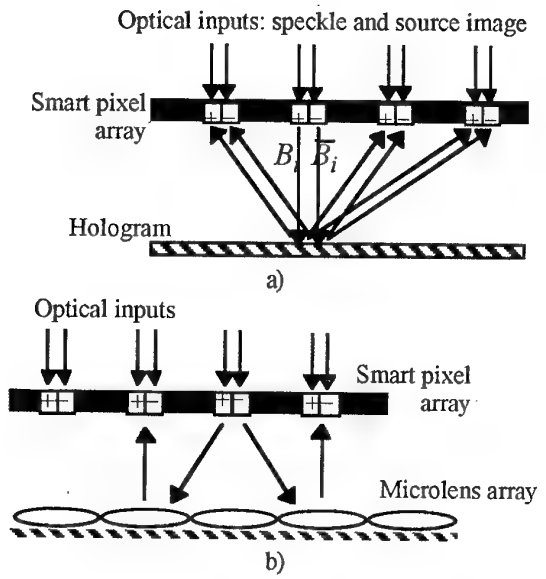


Figure 11 Possible implementations of the interconnection patterns described. (a) Diffractive optics can produce the pattern; (b) refractive optics can also be used.

Many different optoelectronic devices could conceivably be used in the elaboration of this technique: SEEDs, optically activated VCSEL, etc. We used PnpN photothyristors in a preliminary demonstration. Combined in a differential pair, two PnpN photothyristors act as a comparator [25] and allow cascaded operations. The operating frequency can be in the range of 100 kHz. It is limited by the light power available rather than by the PnpN photothyristor response time. With the utilization of PnpN photothyristors, a high pixel density is possible ($> 10^5$ elements per cm^2) and the system design is simple. We have demonstrated this behavior of thresholding detector for stochastic algorithms with a PnpN array.

One problem is that PnpN photothyristors emit as LEDs rather than as lasers : their emission diagram covers a large solid angle (about 2π), which is a cause of waste of light. To overcome this deficiency, an array of microlenses could be adjusted atop of the smart pixel array to concentrate the light emitted by every element. The array can be added to both proposed architectures and would give result in better performances.

7 - CONCLUSION

Our intention with this chapter was to emphasize the possible use of the clearest asset of optics in future computing systems, i.e. the large potential number of interconnects. It obviously applies to cases where the number of interconnects needed is largest, i.e. in systems with the largest degree of parallelism : this include "massively parallel" electronic computing systems, with present goals in the range of thousands of powerful, general purpose processors interacting on common tasks, and "massively parallel" optoelectronic cellular automata, with optical scale parallelism, i.e. at least tens of thousands of small specialized processing elements performing simple vision tasks in video real time. Chief among the challenges faced by the development of optical computing in these domains is system integration, with the need of cheap and reliable passive microoptic fabrication technologies and of packaging technologies maximally compatible with those of present electronic systems.

REFERENCES

1 - J. Jewell, AT&T Bell Labs, 1989.

2 - Optics and Photonics News, février 1993 ; E. Zeeb et al., Optical Computing '94, Edinburgh, August 1994, Proceedings, Institute of Physics Conference Series 139 (1995).

3 - Suitable journals on this question include Appl. Phys. Let. et J. Lightwave Technol..

4 - P. Koppa et al., J. Phys III (France) 4 2405-2411 (1994).

5 - T. Rivera et al., Appl. Phys. Let. 64 869-871 (1994).

6 - D.A.B. Miller et al., Appl. Phys. Let. 45 13-15 (1984).

7 - D.A.B. Miller, Optical Computing '94, Edinburgh, August 1994.

8 - J. Pankove et al., "a pnpn optical switch", Proc. SPIE 963, 191-197, 1988 ; K. Kasahara et al. Appl. Phys. Lett. 52 679-680 (1988).

9 - P. Heremans, M. Kuijk, Optical Computing '94, Edinburgh, August 1994

10 - See for example the proceedings of the French workshop on Micro-Optical Processes and Models, Metz, April 1995, and the associated special issue of the journal Entropie, to appear.

11 - H. Dammann, K. Görtler, Opt. Commun. 3 312-316 (1971) ; J.L. Tribillon, J.E.O.S. A, Pure Appl. Opt. 3, 389-411 (1994).

12 - J.W. Parker, Optical Computing '90, Kyoto, avril 1990 ; Th. Lemoine, 10ème journée d'étude d'opto-informatique SEE/SFO, Gif/Yvette, novembre 1994.

13 - See for example Javidi B. and Réfrégier P., editors, Optical Pattern Recognition, Euro-American Workshop (Bellingham, Washington : SPIE Press), or Horner J.L., and Javidi B., editors, Special section on Pattern Recognition, Opt. Engin. 33, 1751-1862 (1994).

14 - Rajbenbach H. et al., "Compact photorefractive correlator for robotic applications," Appl. Opt. 31, 5666-5674 (1993).

15 - Taboury J. et al., "Optical cellular processor architectures," part I, Appl. Opt. 27 1643-1650 (1988), part II, Appl. Opt. 28 3138-3147 (1989). Chavel P. and Taboury J., "Binary optical cellular automata : concepts and architectures," *SPIE Proc.*

CR35 245-265.

16 - Huang A., "Parallel algorithms for optical digital computers", IEEE 10th International Optical Computing Conference, 13-17 (1983). Huang K.S., Jenkins B.K., Sawchuk A.A., "Binary image algebra and digital optical cellular image processor design," Computer Vis. Graph. Image Proc. 45 295-345 (1989).

17 - Bouthemy P. and Lalande P., "Recovery of moving object masks in an image sequence using local spatiotemporal contextual information," Opt. Engin. 32 1205-1212 (1993).

18 - Geman S and Geman D, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", IEEE Trans Pattern Anal. Mach. Intel. PAMI6 721-741 (1984).

19 - J. Marroquin, S.Mitter, T. Poggio, "Probabilistic Solution of Ill-posed Problems in Computational Vision", J. Am. Stat. Assoc., 82, 397, 1982

20 - Lalanne Ph., "Progress towards the optoelectronic implementation of stochastic artificial retinae," submitted for publication.

21 - Goodman J.W., "Statistical properties of laser speckle patterns," in "Laser speckle and related phenomena", J.C. Dainty, editor (Springer Verlag, Berlin), 9-74 (1975).

22 - Lalanne Ph et al., "Optoelectronic devices for Boltzmann machines and simulated annealing," Opt. Engin. 32 1904-1914 (1993).

23 - Prémont G. et al, "Optical thyristor based stochastic elementary processor," Optical Computing 94, Proceedings, Institute of Physics Conference Series 139, 67-70 (1995).

24 - P. Lalanne, G. Prémont, D. Prévost, P. Chavel, "Stochastic Optoelectronic Retinae for Vision Tasks", Optical Computing '94, Edinburgh, August 1994, Proceedings, Institute of Physics Conference Series 139, 295-313 (1995).

25 - P. Heremans, M. Kuijk, R. Vounckx, G. Borghs, "Fast and Sensitive Two-terminal Double-heterojunction Optical Thyristors", Microelectronics Eng., 19 49 (1992).

Parallel Accessed Optical Storage

Sadik Esener
Electrical and Computer Engineering Department, 0407
University of California San Diego, La Jolla, CA 92093
(619) 534-2732

ABSTRACT

The computational power of current high-performance computers is increasingly limited by data storage and retrieval rates rather than the processing power of the central processing units. No single existing memory technology can combine the required fast access and large data capacity. Instead, a hierarchy of serial access memory devices has provided a performance continuum which allows a balanced system design. Conventional memory technology can only marginally support the needs of high performance computers in terms of required capacity, data rates, access times and cost. Significant gaps in secondary and tertiary storage have emerged which make storage hierarchy design increasingly difficult. This paper reviews a radically different approach to data storage using the parallelism and three dimensionality of optical storage. 3-D optical storage has the potential to significantly alter the present hierarchy and fill the pressing need for high performance secondary and tertiary storage systems.

1. INTRODUCTION

For many applications, the processing speed of today's high performance computers is increasingly limited by the data storage and retrieval rates as well as capacity of current memory systems rather than by the processing power of the central processing units. During the past fifty years many memory technologies have been developed. Despite intense competition, several widely different approaches are in current use, including magnetic and optical tape and disks (hard disks, floppies, and disk stacks)[1], and electronic static (SRAM)[2] and dynamic (DRAM)[3] random-access memory. This proliferation of technologies exists because each technology has different strengths and weaknesses in terms of its capacity, access time, data transfer rate, storage persistence time and cost per megabyte. No single technology can achieve maximum performance in all these characteristics at once. Instead, modern computing systems use a *hierarchy* of memories rather than a single type.[4] The memory hierarchy approach utilizes the strong points of each technology to create an effective memory system that maximizes overall computer

Paper presented at the AGARD SPP Lecture Series on "Optical Processing and Computing" held in Paris, France from 12-13 October 1995; Rome, Italy from 16-17 October 1995 and Ankara, Turkey from 19-20 October 1995 and published in LS-199.

performance/cost.

Volume holographic storage concepts were first put forward in the sixties. Early attempts in optical archival storage were unable to compete with the rapidly improving electronic storage technologies.[5] Until recently, computer technology was incapable of making full use of a major advantage of optical volume storage: the high data rate possible with massively parallel access.

In this paper, we briefly describe the present challenges in designing memory systems with currently available data storage systems and the performance required by future memory systems. We then explore means by which optical storage systems may meet these requirements. We first discuss the present capabilities of optical disk systems and evaluate the potential benefits and means of developing higher density and parallel accessed optical storage systems. We then focus on 3-D optical storage, beginning with the underlying fundamentals and then moving to the various storage materials. We illustrate aspects of system design using one particular approach- two photon storage. Finally, we extract the performance potentials of this system and show how it can alleviate some of the problems currently encountered in hierarchical memory system design.

2. A REVIEW OF PRESENT STORAGE HIERARCHY

In standard sequential computer architecture there are four major levels of the storage hierarchy: cache, main, secondary and tertiary (archival) (see Figure 1). In parallel computers, the difference between main and cache storage becomes less distinct and they are jointly called primary memory. Primary memories are currently implemented in silicon: cache memory as local storage within the processing chip and main memory as RAM and DRAM chips located on the same board. The access

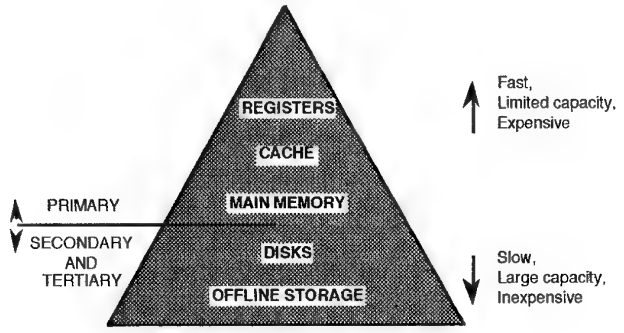


Figure 1: Memory hierarchy in existing computing architecture.

times of primary memories are comparable to a 10 ns clock cycle, but their

data capacity is limited (10-100MByte for main), although it has been doubling every year. Secondary memories, such as magnetic or optical disk drives, have significantly increased capacity (10GByte) with significantly lower cost per MByte. However, the access times are on the order of 10 ms, and this access time is for many computing applications presently the major performance limiting factor. Archival (or tertiary) storage holds huge amounts of data (Terabytes), however the time to access the data is on the order of minutes to hours. Presently, archival data storage systems require large installations based on disk farms and tapes, often operated off-line. Archival storage does not necessarily require many write operations and write-once read many (WORM) systems are acceptable. Despite having the lowest cost per Mbyte, archival storage is typically the most expensive single component of modern supercomputer installations. Special computing applications such as image processing may also use a buffer memory for fast parallel data acquisition before downloading into more permanent storage. Buffer memories tend to be quite application specific, and will not be discussed in this paper.

In Figures 2, 3 and 4 we show the current memory technologies in terms of their critical characteristics. As one can observe from Fig. 2 there is a four order of magnitude performance gap in access time between electronic RAMs and secondary storage devices such as disks. The width of this gap is doubling each year, forcing the development of new secondary storage systems. In addition, because the processing power doubles every year and the memory density only every two and half years, an increasingly large gap also exists in terms of memory capacity. A similar situation exists between secondary storage systems and archival systems (such as tapes) that suffer from slow access times. In addition, Fig. 3 shows that the data rates of current secondary storage technologies are significantly lower than required for full use of parallel processors. Techniques to enhance secondary storage data rates are therefore crucial. Finally, Fig. 4 shows that existing storage systems are too costly for wide use in applications such as databases that require high capacity and fast access times.

Existing serial memory technologies using planar storage media, including electronic RAM, magnetic disk, and optical disk storage, are inadequate to bridge the gaps described in Fig. 2. The data capacity is fundamentally limited by their two-dimensional nature to the storage area divided by the minimum bit size. Similarly, as the storage area increases for higher capacity so does the access time. Another drawback is that the data transfer rates are limited by their sequential nature to the I/O channel bandwidth.

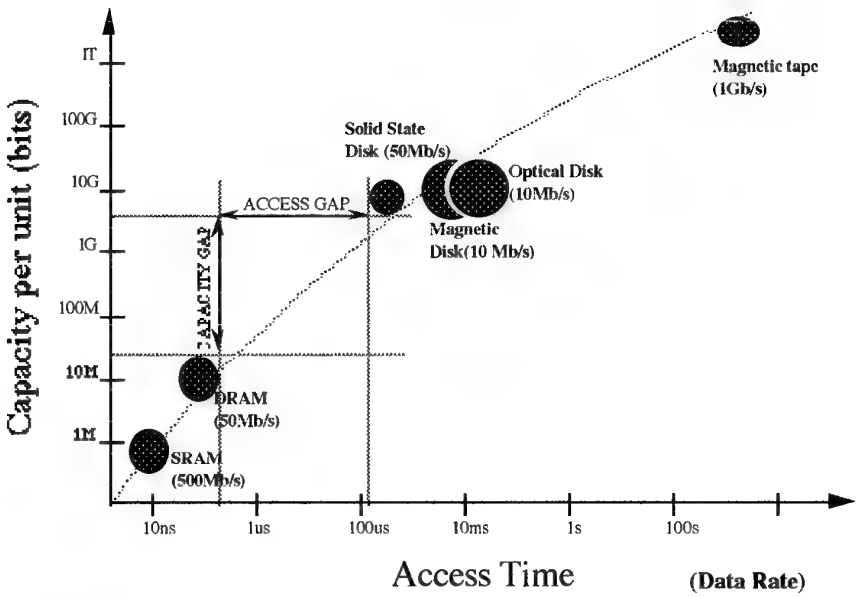


Figure 2: Access speed and capacity of existing memory technologies.

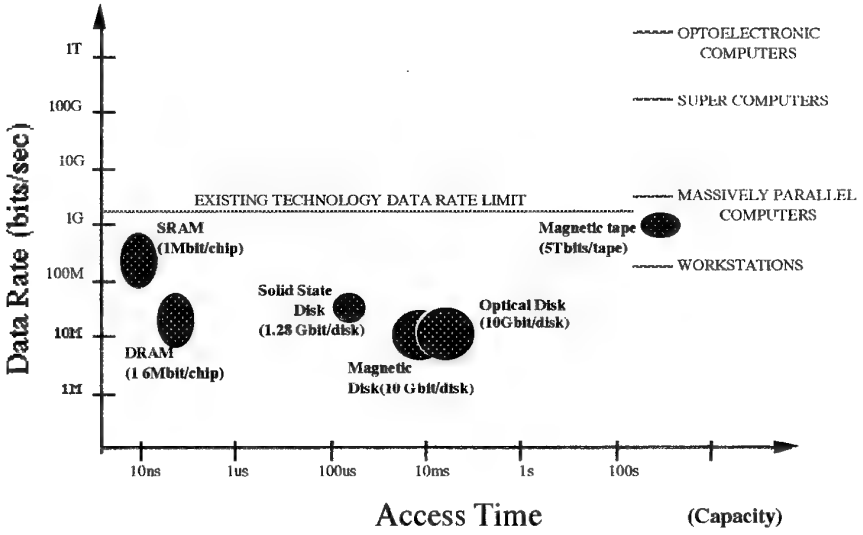


Figure 3: Data rates of existing primary storage technologies.

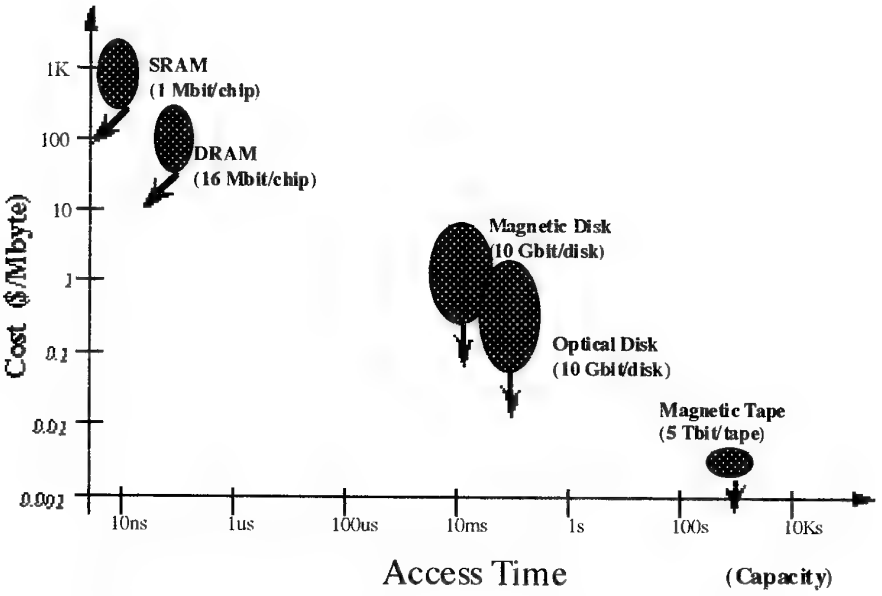


Figure 4: Cost per Mbyte of existing primary storage technologies.

In the following section, we review the current limitations of optical disk storage and explore how they can be modified to increase data density and how higher data rates can be achieved by exploiting parallelism. In the subsequent sections, we will discuss how three-dimensional memories (3DM) surmount the capacity and access time limitations by extending the storage into the third dimension.

3. OPTICAL DISK STORAGE: PRESENT AND FUTURE

Optical disks have become a major component in many computer systems due to their high capacity, low cost, data integrity and security (due to removability). In this section, we first describe the current status of optical disk technology. Then, we present the future projections for both serial and parallel access of the data in optical disk systems.

3.1 Performance of current serial optical disk systems

Present optical disk storage systems provide about 500Mb/in^2 storage density with a total maximum capacity of about 10GBytes for 14" disks. The data density is limited through diffraction by the wavelength (780-830 nm) at which data is currently recorded on optical disks. Research on short wavelength integrated lasers and on the use of superresolution in the optical read/write channel is presently being conducted to reduce the physical bit size and reach data storage densities approaching 10Gb/in^2. Simultaneously, research is being carried out to develop the necessary higher resolution optical disk media capable of read-write-erase operation.

The main disadvantages of commercially available optical disk systems are their low data rate and poor access time performance. Low data rates are due to the relatively slow disk rotation speeds (<3600 rpm) which are mainly limited by the tracking and focusing servo loops. The slow access times are mainly due to the weight of the heads and the slow rotation speeds that are respectively responsible for large seek times and large latency times. The detection speed is governed by the limited source power, the detector sensitivity, and the write and damage thresholds of the disk material. Present development efforts are aiming towards low mass optical heads which reduce the response time and increase the allowed disk rotation speed. These new heads incorporate holographic or integrated optical elements with substantially less total weight than their bulk optic counterparts. In addition, efficient error correction codes and higher power lasers (e.g. array lasers with combined beams) are used to increase the detection rates and avoid limiting the maximum data rate. Solutions derived from magnetic disk technology have also been envisioned (e.g. flying heads or air bearing drives) but have not been pursued since their adoption would eliminate one of the main advantages of optical disks: removability.

It is expected that these developments will result in a continued improvement of the optical disk performance. However, the resulting performance increase is not expected to satisfy the ever rising demands of future computers which will require terabits of data capacity accessed in sub-milliseconds with data rates approaching Tb/s. Therefore, optical disk storage research must seek alternative approaches such as the use of parallel data access to substantially improve upon the current data rates.

3.2. Parallel access optical disk systems

A major improvement in the data transfer rates of optical disks can be achieved by accessing the data in parallel. The concept was first proposed in 1980[6] and later by several others.[7,8,9,10] The research efforts are divided in two main lines of work: one approach uses either a single multiple beam head or multiple independent integrated laser heads to read several tracks in parallel, while the other consists of reading coded data blocks with a single laser beam. In this latter case, both imaging[3] and Fourier transform[2] systems have been proposed. In the following, we describe a hybrid approach currently under development at UCSD.[4]

The system is designed to read 1-D data blocks distributed radially on the disk's active surface and generate 2-D output binary images. It has the unique advantage that no mechanical motion of the head is required for data addressing, focusing or tracking. This allows a data rate of up to 1 GByte/sec for standard disk rotation speeds of 2400 rpm.

The data blocks are stored on the disk as 1-D Fourier-transform

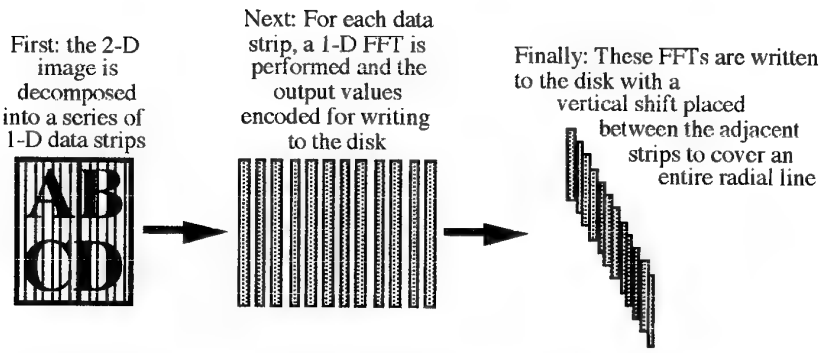


Figure 5: Image encoding in UCSD's parallel readout optical disk system.

computer generated holograms (CGH). Each one of them is calculated to reconstruct one column of a 2-D output image. During the sequential recording process, all the CGH for a given 2-D image are laid-out side by side and radially shifted from one another until they span the entire radius of the disk active surface as shown in Figure 5. During the parallel readout process, all of the data blocks of a single image are illuminated at once. Data addressing is achieved solely through the disk rotation so that the entire memory can be read in one rotation. The Fourier direction of the holograms is along the radial direction of the disk. Thus, due to the shift invariance properties of Fourier transform holograms, the tracking servo requirement can be eliminated. Under-illuminated Fourier Transform holograms will still reconstruct, although with a loss in output signal-to-noise ratio. Therefore, by slightly under illuminating the hologram block the focus servo requirement can be virtually eliminated.

The optical readout system is shown in Figure 6. It consists solely of three stationary cylindrical lenses. The first illuminates an area on the disk that corresponds to the data blocks of a single output 2-D image. The other two lenses image the data blocks of the disk along the tangential direction and perform a Fourier transform along the radial direction. In the actual system, these two lenses have been replaced by a single hybrid diffractive-refractive optical element for easier system alignment, reduced aberrations, and improved resolution. A similar optical parallel read out system can also be applied to increase the data rates of optical tapes.

A scaled-down prototype using a commercially available plastic 5.25" WORM disk[11] has been implemented that reads out 16x16 pixel images at a maximum rotation speed of 30 rpm. The complete characteristics of the prototype system have been reported.[12] Experiments performed on the prototype system showed that a full scale system could operate with a data transfer rate of 10GBytes/sec and a disk capacity of 300MBytes.

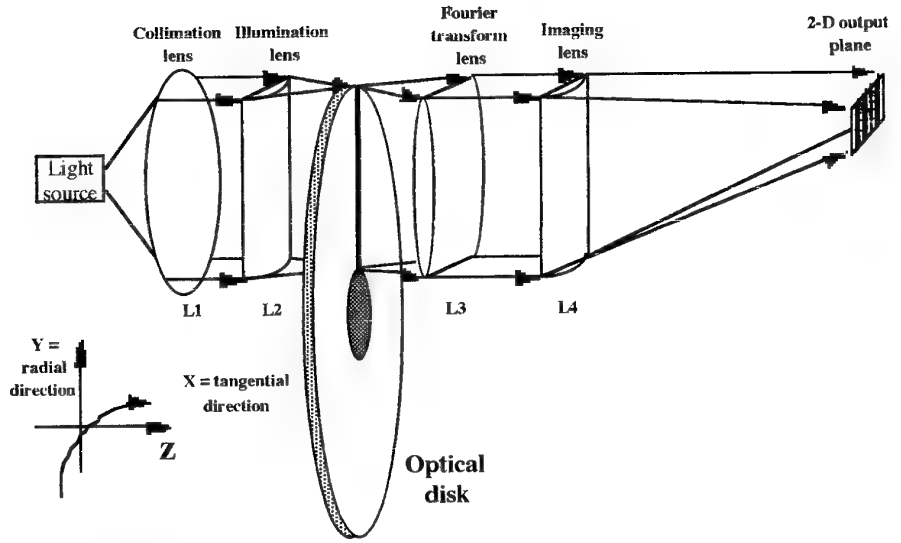


Figure 6: Parallel readout optical disk system. Cylindrical lenses combine imaging and Fourier transforming to reconstruct the recorded image.

These results indicate the optical storage systems with parallel read-out capabilities may shortly enter the realm of practice. Unfortunately, by exploiting parallelism for higher data rates the storage capacity of the system is reduced. Therefore, parallel accessed two dimensional storage systems should be useful primarily when the operations to be performed on the data can be parallelized. For example, an associative memory system based on the parallel accessed optical disk system described above is currently being constructed at UCSD.[12]

To become useful on a larger span of applications, however, optical storage systems must provide higher capacity for faster data rates and access times. This can be achieved by employing 3-D optical storage systems.

4. PHYSICAL FUNDAMENTALS OF 3-D OPTICAL STORAGE

We will briefly consider the fundamental principles underlying optical storage and retrieval of information from three dimensional media. These concerns are largely independent of the particular materials used, which will be discussed in the Section 5.

4.1 Definition and classification of 3-D optical memories

Present optical memories store one-dimensional information sequences in a two-dimensional space using two spatial coordinates (either rectangular {x-location, y-location} or polar {radius, angle}) to assign a

precise physical location to each bit. The maximum storage capacity is given by A/λ^2, where A is the storage area and λ is the optical wavelength. A 3-D memory can be defined as a single memory unit where three independent coordinates are used to specify the address of the information. These coordinates may be entirely spatial {x-location, y-location, z-location} but may also use other physical parameters (e.g., {radius, angle, wavelength}). In 3-D optical memories, information can be partitioned into binary bit-planes (or images) that are stacked in the third dimension. One memory operation (write, read, or erase) is performed on the entire plane of bits, thus achieving 2-D parallel access of the data in a single cycle of operation. As in the parallel access optical disk system described in Section 3, this results in a significant increase in the maximum achievable data transfer rate over conventional systems. Since, in addition, optical 3D memories can achieve very high density (10^{11}-10^{12} bits/cm^3) they offer the potential for larger storage capacity systems with lower access times and higher data transfer rates than conventional 2-D memories.

3-D memories can be classified as using either bit-oriented (localized) or holographic (distributed) storage, and can be further classified by the type of physical coordinates used to address the data, as is shown in Figure 7. The physical mechanisms which can be used for recording and addressing in 3-D optical memories are discussed in Section 4.3. First, however, we discuss the characteristics of bit-oriented and holographic data storage formats.

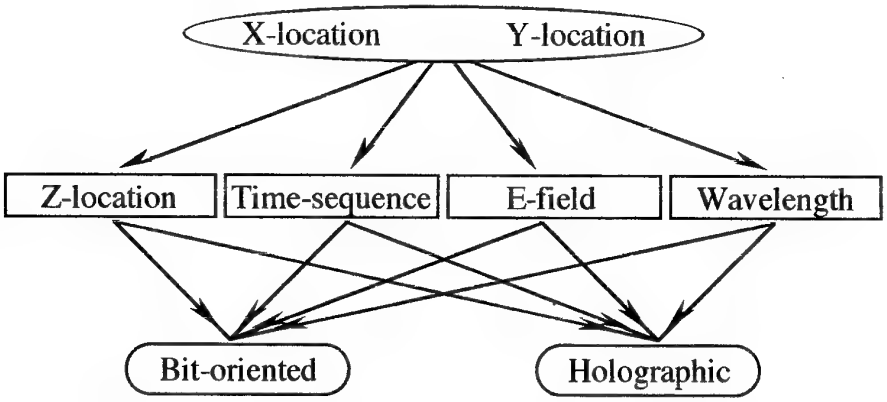


Figure 7: Classification of 3-D optical memories. Two spatial dimensions are used for the 2-D images. A third dimension is defined by an additional physical addressing parameter. The information can then be stored in either bit-oriented or holographic format.

4. 2 Physical principles of 3-D information access

4.2.1 Bit-oriented(local) storage

Probably the most direct approach to volume data storage is to partition a storage volume into sub-volumes, each containing one datum. The traditional 3-D storage technology --books-- uses this approach. The location of each letter in a book can be described by its x-y coordinates on the page and its depth in the media (its page number). If you imagine a book printed on transparent pages, permanently stuck together, then the concerns associated with optical storage and readout become obvious.

Two dimensional optical wavefronts can access data in three spatial dimensions only by involving an additional mechanism. Otherwise, the image simply propagates through the storage medium, superimposing the effects of all the stored pages. The selection of a particular position in the third dimension comes from an additional parameter. In two photon materials, the additional requirement is that photons of a second optical wavelength must be present in order for the first beam to interact with the material. In spectral hole burning, the additional parameter is the wavelength. And in photon echo or free-space storage, the additional addressing parameter is time.

Diffraction effects enforce a theoretical limit to the maximum achievable spatial resolution of any optical memory. The minimum spot size is conventionally given as $\delta = 1.22\lambda F$, where λ is the optical wavelength and F is the f-number of the system (focal length divided by clear aperture).[13] The resolution in the third dimension depends on the type of addressing parameter used, as discussed in Section 4.3.

4.2.2 Holographic (distributed) storage

In bit-oriented memories the data resides in a restricted region. This means that if any portion of the storage media is damaged or blocked, the data which was stored in that region is lost, but the other data is completely unaffected. In more concrete terms, if half of your material is destroyed, half of your data is lost. This is not the case for holographic storage, where the information about each stored bit is distributed throughout a large region. If a portion of the storage media is damaged or blocked, instead of causing catastrophic loss of some of the data, all of the data is partially degraded. If half of the material is destroyed, all of the data remains to some degree. Whether *any* of the data is still legible depends on the extent of the damage. For common types of damage, such as surface dust or smudges, holograms are remarkably robust. This has generated interest in holographic data storage, and despite the more

complex optical systems and the coherent sources required there has been continued research in the field since the early 1960s[5,14]. Excellent texts on holography are available.[15,16,17] We will only provide a brief description of hologram characteristics as they relate to data storage.

Holograms are created by recording the interference pattern of two optical wavefronts. The storage media can record the fringes as index and/or amplitude modulation. When the recording is illuminated by one of the wavefronts (the reference beam), the other wavefront (the object beam) is reproduced. Planar holograms, which record only a 2-D interference pattern, have quite different behavior from volume holograms, which also record the variation through a depth significantly greater than the fringe spacing. Planar holograms are theoretically limited to 6.25% efficiency if recorded in amplitude, and 33.9% efficiency if recorded in phase.[15] The maximum diffraction efficiency of a volume absorption hologram is theoretically limited to less than 7.2%. Pure phase volume holograms, however, can approach 100% efficiency.

The most important difference between planar and volume holograms is their selectivity with respect to the readout (reference) wavefront. When the reference wavefront is tilted, planar holograms still diffract with nearly full efficiency, producing a reconstruction tilted by a similar angle from the original. For a volume hologram to diffract efficiently, however, the original recording wavefront must be closely reproduced. This effect, called Bragg selectivity, allows a volume hologram to independently store and recall multiple superimposed holograms provided that each has a sufficiently distinct reference beam. The detuning angle to drop to the first zero of diffracted intensity is approximately Λ/T radians, where Λ is the grating wavelength and T is the hologram thickness.[15] For a 1 cm thick hologram, this angle is on the order of hundredths of a degree. Similarly, a volume hologram can have a 1 Å wavelength selectivity.[18]

If an additional physical parameter is used for addressing, such as time sequence, a 3-D optical memory can be constructed using planar holograms. The following discussion applies only to volume hologram multiplexing, as for example in photorefractive crystals.

The theoretical upper limit of the storage capacity of a volume hologram[19] is given by V/λ^3. Constraints arising from optical system[20] and storage media characteristics[21] significantly reduce the achievable volume holographic storage density. However, storage densities in excess of 10^9 b/cm^3 can realistically be achieved.[22] The information which can be extracted from the volume at a particular instant is still limited by the hologram aperture to A/λ^2 for each wavelength used for readout. The large theoretical capacity must be accessed in pages, multiplexed in the volume by wavelength or phase.

The goal in multiplexing holograms is to maximize storage capacity (the number of images, their resolution, and their diffraction efficiency) while minimizing crosstalk. Volume holograms can be multiplexed by wavelength and phase (and polarization in some materials), and of course by the volume illuminated. Phase multiplexing includes reference beam tilt or curvature as well as more complex (or even random) phase patterns. Polarization multiplexing is not used for data storage since there are only two orthogonal polarizations. However, volume and phase multiplexing have been extensively studied for storage applications.[23,24,25] More recently, with the availability of effective color tunable lasers wavelength multiplexing has been a topic of investigation, especially in reflection volume holograms, which are highly color selective.[18]

Crosstalk in volume holograms arises from two major sources, Bragg degeneracy and higher order diffraction. Bragg degeneracy comes from the fact that in addition to the exact reference beam angle, there is a cone of vectors which match the Bragg condition. Similarly, if the readout wavelength (or angle) is changed, a simple plane–wave hologram can still be read at full efficiency by using an appropriate angle (or wavelength) which matches the Bragg condition. This behavior effectively vanishes as the hologram becomes more complex, because the Bragg condition can only be satisfied simultaneously for the entire angular spectra by the original reference wavefront. Scattering from locally matched portions of the grating can only produce background noise, whose intensity becomes negligible as the hologram thickness increases. In currently available materials, the storage capacity seems to be limited more by the material dynamic range rather than by crosstalk resulting from multiplexing.

4.3 Physical mechanisms of 3-D optical storage

Optical recording can occur only when photons are absorbed by matter. If this absorption changes any of the material's optical properties, then data can be recorded and read out optically. The basic characteristics of an optical wavefront are its wavelength, amplitude (intensity), polarization, and phase (the direction of a beam is determined by the wavefront's phase profile). Any optical storage media affects one or more of these characteristics. For example, developed photographic film becomes opaque where it was illuminated. Other materials can change their index of refraction (e.g. photorefractive crystals and dichromated gelatin film), their surface reflectance (write-once optical disks), their absorption spectra (spectral hole burning materials), their fluorescence spectra (photochromic materials), etc., in response to optical illumination.

The 'location' of stored data is given by the set of physical parameters required to retrieve that data. This may the three spatial coordinates in a

volume, but might also include the readout wavelength and polarization, an external applied voltage, an external magnetic field, and even the timing or sequence with which these parameters are applied. A set of address coordinates or dimensions is said to be orthogonal if they are totally independent of each other, as are, for example, the three spatial coordinates. The effective extent of each dimension can be defined as the range over which storage is possible divided by the resolution (the smallest resolvable division) maintained over that range. These considerations may be limited by material or experimental constraints, or by more basic principles. Polarization is a dimension whose extent is only the two orthogonal polarizations. The extent of the horizontal and vertical spatial dimensions of an image can be huge, however, with a range of centimeters and a resolution of microns.

In 3-D memories, the coordinates that specify the location of the information can be spatial, spectral, or temporal, giving rise to a variety of 3-D memory concepts using different materials with various properties. For example, 2-photon absorption materials can form a true volume memory while spectral hole burning materials provide a 3-D storage medium with two spatial and one spectral dimension. Optical memories of even higher dimension (4-D or 5-D memories) are possible provided that appropriate materials are identified and the necessary systems are developed. The image bearing character of light and the planar nature of detector arrays make three the magic number for optical memories, since this allows a time sequence of 2-D images to be retrieved, detected and processed in parallel.

Multiplexing with parameters such as wavelength or time-sequence is in effect a way of volume multiplexing at sub-wavelength scales. A cubic micron is the smallest optically resolvable volume. However, there are trillions of distinct atoms within that volume and each may have a different optical response. For example, spectral hole burning materials use the fact that similar molecules within a bulk material have different absorption bands due to variations in local stresses.[46] There may be billions of molecules with each characteristic absorption spectrum within a single cubic micron, so that each different wavelength seems to illuminate untouched material. Temperatures near absolute zero help prevent transfer of information between adjacent molecules. However, the material sensitivity and readout efficiency are probably not as high as they might have been if the material was composed of a single pure species with uniform response. The difference between simple volume multiplexing and these more exotic alternative addressing techniques is that only in volume multiplexing at optically resolvable scales can *all* of the local material be entirely devoted to recording a single bit of information.

5. 3-D OPTICAL STORAGE MATERIALS

5.1 Photorefractives

Photorefractive crystals (PRC) are a reversible phase storage media which have been extensively studied for holographic storage applications since their discovery in 1965. The sensitivity approaches that of photographic emulsions, and the index modulation is large enough to record nearly 100% efficient holograms. Crystal sizes range up to many cm^3, depending on the particular material. PRC have three characteristics which combine to make them different from any other holographic recording media: 1) in–situ recording and development, 2) energy coupling between recording beams, and 3) write–erase capability with infinite cycling. The basic mechanisms of the effect have been uncovered and analyzed, and a number of excellent overviews of photorefractive processes and materials are available.[26,27,28] The process can be qualitatively described as follows.

Consider two plane waves crossing in a photoconductive and electro-optic material. They produce a sinusoidal fringe pattern I(z) (see Figure 8). In regions of high intensity, photoabsorption excites charge carriers to the conduction band, allowing them to drift or diffuse towards darker areas, where they are trapped. The resultant space charge distribution generates an electric field E(z). This field modulates the refractive index by the electro-optic effect, creating a sinusoidal phase grating $\Delta n(z)$. If the writing beams are cut, the phase grating remains stored in the material until thermal or optical excitation redistributes the trapped charge carriers. If the crystal is uniformly illuminated, the index modulation is erased by the same process. The cycle can be infinitely repeated.

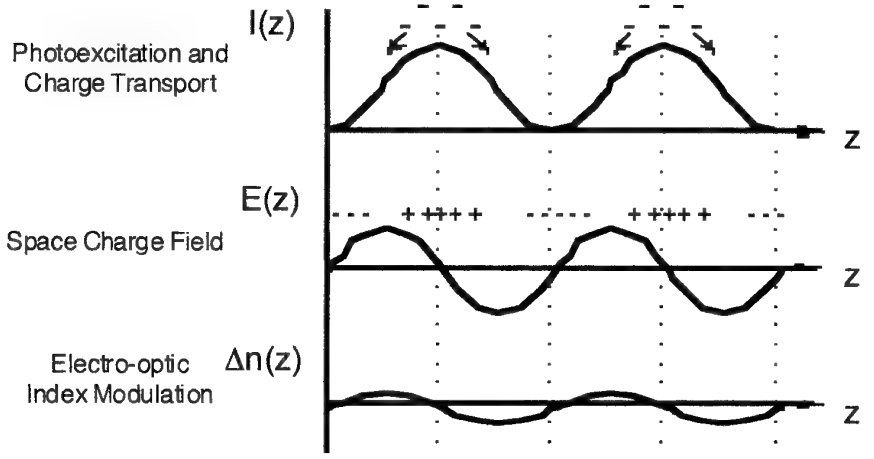


Figure 8: The photorefractive effect.

A large number of photorefractive materials have been identified. They can be broken roughly into three groups; ferroelectric oxides, cubic oxides (sillenides), and semi–insulating compound semiconductors. For storage applications materials with long relaxation times are preferable, making ferroelectric oxides the material of choice. They include Lithium Niobate[29], Potassium[30] and Potassium-Tantalum[31] Niobate, Barium Titanate[32], and Strontium Barium Niobate.[33] Their large electro–optic coefficients produce a large saturation index modulation. Their low photoconductivities ensure long dark storage lifetimes, but also result in lower sensitivity (the index change per unit absorbed optical energy per unit area). For example, the sensitivity of $BaTiO_3$ is about 10^{-3} cm^2/J, producing a response time of 0.1–1 second for 1 W/cm^2 input power. Another photorefractive with potential storage applications is ceramic Lead–Lanthanum-Zirconium–Titanate[34]

Photorefractive holograms in ferroelectric oxides are of high optical quality, and have been shown capable of reconstructing images of 10^6 bits or more.[28] As many as 5000 angle multiplexed images have been recorded in a single $LiNbO_3$ crystal.[35] The fundamental limitations on PRC sensitivity[36] come from the quantum efficiency of the photoabsorption, the carrier lifetime and mobility, and the electro–optic coefficient. The physical resolution is high enough to record even the highest frequency reflection grating accurately, although the strength of the response does depend on grating frequency and orientation. In the absence of an external applied field,[32] the self-developing index modulation response follows exponential write and erase curves, with a response time τ which is inversely proportional to the writing intensity. While the photorefractive effect is generally thought of as slow (on the order of milliseconds or longer), with sufficiently high intensity (GW/cm^2) even picosecond responses are possible,[37] although full saturation index modulation is not achieved. The maximum index modulation depends only on the contrast between the interacting beams, not on the absolute intensity of the interacting light (provided the intensity is above a threshold determined by the spontaneous redistribution of charge carriers).

The fact that the photorefractive process is reversible means that there is erasure during readout of recorded holograms. If the ratio of recording to readout intensities is large enough, the holograms can persist for a long time. The fundamental limit to storage lifetime is the time for the charge distribution to spontaneously relax to a uniform equilibrium, which is determined by the material's dark conductivity. This can range from hours in $BaTiO_3$ up to years for $LiNbO_3$. The photorefractive index grating can be semi–permanently stored, or "fixed", so that the hologram can be read out at high intensity for extended periods without significant erasure.

Fixation has been accomplished by two-photon absorption in $LiNbO_3$, KTN, and PLZT, or by ionic charge compensation, where the electron distribution is converted into a similar distribution of non–mobile ions. Permanent fixation by charge compensation has been demonstrated in $LiNbO_3$.[38] However, the materials and mechanisms of photorefractive fixation are not fully understood, and some of the early impressive results have proven difficult to reproduce.

When holograms are multiplexed in a photorefractive crystal, each exposure tends to partially erase the existing information in the crystal. To superimpose holograms and end with a specified relative diffraction efficiency - usually equal efficiency for all holograms - requires some kind of recording timetable. In scheduled recording,[22] the first hologram is recorded to the maximum achievable efficiency. The exposures for each subsequent hologram are shorter and shorter, so that at the end a set of equal efficiency holograms is obtained. The recording timetable is calculated from the material's index response curve. An alternative approach, called incremental recording,[24] records each hologram with a series of exposures which are very short compared to the crystal's response time. During recording, each image and reference pair is sequentially displayed, repetitively cycling through all the images until the process reaches saturation. The final diffraction efficiency of the multiplexed holograms decreases as the number of superimposed holograms grows. The diffraction efficiency η of N superimposed volume phase holograms is approximately $\sin^2(\pi\Delta nT/N\lambda\cos\theta)$, where Δn is the index modulation, T is the thickness of the crystal, λ is the optical wavelength, and θ is the internal angle between the recording beams. The writing and erasing processes have been assumed to be exponential with the same time constant. For large N, the efficiency drops as the square of N. Increasing T increases both hologram efficiency and selectivity, making thick crystals desirable for data storage applications.

Although significant progress has been made at the system level, the characteristics of existing PRC remain far from satisfactory for reliable low cost systems. Major improvements in crystal dimensions, cost, and dynamic range (modulation of the index of refraction) need to be realized before such systems can become competitive. Recently, photorefractive effect has been observed in low cost materials such as plastics and this direction may bring some hope of reducing the cost of PRC materials.

5.2 Irreversible volume holographic materials

Many other materials have been used to record multiplexed volume holograms. Dichromated gelatin[39] can record a permanent index hologram with high efficiency (approaching 100%), but the thickness is limited to

around 100 μm. Photochromic materials, which produce a reversible color change record in response to light, can record an absorption hologram. However, the maximum diffraction efficiency of an absorption hologram is theoretically limited to 7.2%.[15] Photopolymers[40,41] show promise as a high–efficiency moderately thick (25-50 μm) recording media, particularly since they can be developed in–situ (under certain conditions) without wet processing. However, their diffraction efficiency is a strong function of the fringe spacing, and the fringes are sometimes a combination of phase and surface relief modulation.

5.3 Bacteriorodhopsin

A biological material, the bacteriorhodopsin which is a photochromic protein may present an inexpensive alternative to PRC and other more conventional volume holographic materials. The photochemistry associated with bacteriorhodopsin is well documented and the reader is referred to references [42,43, for example] for an in depth treatment. The bacteriorhodopsin has a light absorbing chromophore which is bound to the protein through a protonated Schiff base linkage. Any change in the electronic environment of the binding site of the chromophore results in a change in the spectral characteristics of the overall protein. Such changes can be induced by light absorption at proper wavelengths and temperatures resulting in a photocycle. The photocycle of bacteriorhodopsin is comprised of at least five thermal intermediate states. It has been suggested that at least three of these intermediate states (**bR**, **K** and **M**) have potentials for optoelectronic device applications. Each key intermediate exhibits a unique absorption spectrum. The initial state, **bR**, is characterized by a large absorption maximum in the yellow region of the visible spectrum. At low temperatures (77°K), through light absorption by the chromophore, a nearly instantaneous shift of electron density affects the chromophore-protein link, resulting in photo-isomerization and leading to the formation of the **K** intermediate state. The **K** intermediate exhibits a large absorption in the red. The isomerization speed is very high (psec). For storage applications, the low temperatures required for this process may complicate system packaging and the **M** intermediate gains importance. The formation of the **M** intermediate occurs at room temperature as a result of a series of protein conformational changes. During this process, the proton of the chromophore is transferred to an amino acid of the protein resulting in a highly blue-shifted absorption spectrum. However, the speed of formation of the **M** intermediate is much slower (0.1 msec) then the **K** intermediate and is a limiting factor in the memory writing process. In addition, under normal biological conditions, the **M** intermediate is not stable and reverts to the **bR** intermediate within

10 msec, seriously limiting the memory persistence time. However, by suspending bacteriorhodopsin in a polymer matrix in the presence of certain chemicals the lifetime of the **M** intermediate can be increased to about 30 minutes. This property, coupled with good quantum yields (0.64) in both directions, high photocyclicity (>10^6 cycles), low cost and room temperature operation may make this protein competitive to PRCs and attractive for some primary storage applications that require only short memory persistence times.

The use of a bacteriorhodopsin doped polymer as a holographic medium was originally proposed by Bunkin and coworkers.[44] Due to the different absorption spectra of the intermediates according to the Kramers-Kronig relationship, both amplitude and phase information associated with an optical 3-D interference pattern can be recorded. In places of constructive interference **bR** is driven to the **M** state and in regions of destructive interference no photochemistry is initiated. More recently, the volume holographic storage properties of bacteriorhodopsin in such polymers have been investigated experimentally by Birge and his collaborators who have recorded volume holograms with reasonable diffraction efficiencies (6%) in polymers hosting the protein.[45] The diffraction efficiency was mainly limited by the absorptive nature of the recorded volume gratings preventing the multiplexing of large numbers of holograms. Since phase only gratings will provided very high diffraction efficiency some of the present research effort seeks to eliminate absorption and to record phase only holograms at suitable wavelengths. If and when such holograms can be permanently recorded in bacteriorhodopsin, this highly light sensitive material may become the holographic recording material of choice.

5.4 Spectral hole burning

In this section, a second type of wavelength multiplexing is described which concentrates on photo induced transformations between the various ground states of a molecule and is generally called persistent spectral hole-burning (PSHB).[46] This technique can be called a true wavelength multiplexing scheme because there is complete independence between the information stored at each wavelength. In contrast, wavelength multiplexing in photorefractive crystals (see section 5.1), can suffer from crosstalk problems between the individual wavelength-multiplexed holograms. This crosstalk is due to the fact that the Bragg gratings created by a hologram recorded at one wavelength is visible by all other optical wavelength although they may not Bragg match the gratings.

The types of molecules that are being investigated for PSHB experiments have a large inhomogeneously broadened absorption band

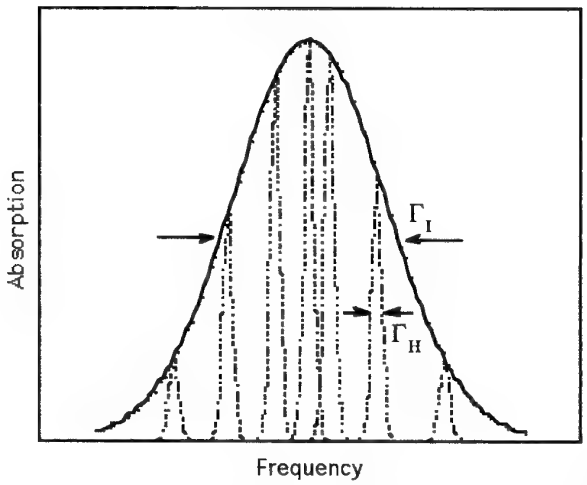
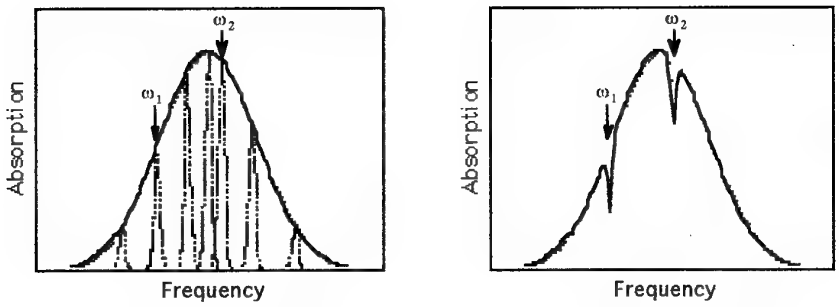


Figure 9. The absorption profile for an idealized PSHB material. The overall linewidth is defined by the inhomogeneously broadened curve of width Γ_I which is composed of a distribution of homogeneously broadened lines of width Γ_H.

comprised of a large number of narrow absorption lines as shown in Figure 9. The solid line shows the overall absorption spectrum of the material as a composite of all of the narrow homogeneously broadened absorption curves (shown as dashed lines).

To record information into the PSHB materials, a tunable wavelength, narrow bandwidth, optical source is used to illuminate the material (where source bandwidth $\Delta\omega << \Gamma_H$). The illuminating light induces many photophysical transformations which dramatically modify the absorption profile near the source energy. Figure 10 shows the absorption profile of a molecule before and after illumination with light frequencies of ω_1 and ω_2. The originally smooth profile has been altered into one with sharp dips near the illuminating frequencies, making it clear why this process has become known as *spectral hole-burning*.

The extent to which the wavelength can be used to store multiple bits of information in a PSHB material is given by the number of distinct holes that can be burned into the absorption profile. This is given by the ratio Γ_I/Γ_H, where Γ_I and Γ_H are the inhomogeneous and homogeneous linewidths, respectively (see Fig. 9). This ratio can be as high as 10^6, but only at extremely low temperatures (< 1K). This is because, in general, the homogeneous linewidth Γ_H increases with temperature[47] and can reach the point where $\Gamma_I \approx \Gamma_H$ for room temperature operation. Recent research efforts have focused on developing new materials that have a large number of spectral holes at room temperature.[48,49]



(a) Before illumination (b) After illumination

Figure 10: The effect of spectral hole burning on the absorption profile. The frequencies of the illuminating sources at ω_1 and ω_2 are indicated with arrows. (a) The smooth absorption profile for an unwritten molecule. (b) The modified absorption profile after illumination.

The original memory system proposed for this technology was a bit-oriented memory that required a thin sheet of PSHB material, a wavelength tunable laser and x-y beam deflectors. Using the two spatial dimensions and one wavelength dimension, a true 3-D memory device was achieved. The spectral holes were burned using an intense beam (write operation) and detected using a much lower laser intensity (read operation). Typically, the presence of a hole is assigned the logical value "1" and the absence of a hole the value "0". One problem with this system design is that it is quite difficult to determine if a spectral hole is present without using a variable threshold detector or other complicated scheme. Recently, absorption holography has been implemented to reduce the background intensity and greatly improve the contrast ratios.[50] Other recent advances have been made to reduce crosstalk by applying an external electric field and sweeping the optical frequency and phase while recording the hologram.[51]

PSHB is an active area of research despite the requirement of low temperature. It is hoped that future advances will raise the operating temperature and provide a storage capacity improvement of three two four orders of magnitude over that of present optical disk systems.

5.5 Photon Echo

This section describes the effect called stimulated photon echo (SPE) and how it can create 3-D memories using time as the third degree of freedom. The coherent transient effect of the SPE was first described by Mossberg as a means of storing/retrieving rapidly varying data in parallel.[52] Figure 11 shows a generalized timing sequence for storing and retrieving a 1-D time-modulated data sequence.

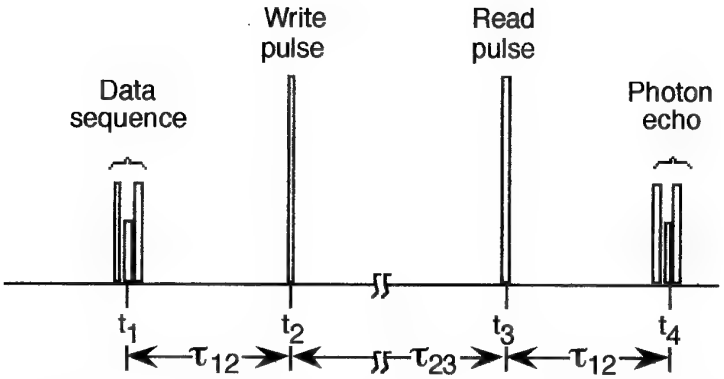


Figure 11: The timing sequences required for the storage and retrieval of time-modulated optical data using stimulated photon echo. The writing process is shown on the left as the data sequence followed by a short write pulse. The data is retrieved by applying a read pulse to stimulate the photon echo that will reconstruct the original data.

The time-modulated data sequence reaches the material at a time t_1 and excites the material according to the optical and temporal frequencies that are present in the signal. A write pulse must be applied to the material, within the homogeneous dephasing time (T_2) of the material, to create a temporal hologram within the ground- and excited- state population densities. This hologram now contains information on the temporal structure as well as the spectral structure of the data sequence. By applying the read pulse at a much later time ($\tau_{23} > T_2$), a photon echo will be emitted by the material with a delay of precisely τ_{12}. The echo can be stimulated multiple times, but each time the readout signal becomes weaker because of the disruption from the energy deposited by the readout process. Eventually, the input signal must be refreshed or the sample will return to the original ground state population density.

There are several conditions that must be satisfied to create a photon echo that will accurately represent the input data. The first requirement is that the material must be very cold (less than a few degrees K) so that all of the excitations and transitions are caused solely by the optical pulses. The second requirement is that the material have an inhomogeneously broadened absorption line (see section 5.4 on persistent spectral hole burning) that is capable of recording both spectral and spatial information.

The types of memories that have been proposed using the SPE are called spatial-temporal memories and usually store information in two spatial dimensions (x, y) and the dimension of time to locate any single memory bit. The extent to which time domain can be used is given by the

amount of temporal data that can be stored/retrieved in a single echo. This value depends on the characteristics of the material in the following way. First, the minimum pulsewidth that can be recorded must be longer that the phase relaxation time T_2^*. Second, the entire data sequence and write pulse must occur in an interval of less than the overall coherence time of the excited state T_2. The ratio of T_2/T_2^* is typically in the range 100-1000.[46]

There are similarities between stimulated photon echo and persistent spectral hole burning in that they use the same types of materials and require very low temperatures. One important difference is that SPE does not require a tunable laser source. It does however require a modulated laser beam and detector that can attain speeds of up to 100 MHz.

5.6 .Multi-wavelength storage materials

Another approach to volume data access is to use the difference generated in the absorption characteristics of written molecules which absorb at different wavelengths. This mechanism is used in the multi-frequency optical volume memory[53] that is being developed in Japan. This memory consists of many layers of J-aggregate photochromic Langmuir-Blodget thin films having sharp and distinct absorption bands as shown in Figure 12. By pre-exposing the films with appropriate UV radiation the required sharp difference in the absorption spectrum of each layer can be synthesized. To write a bit, molecules are excited with a UV beam while the reading of stored bits is performed by selecting the appropriate storage layer using a wavelength tunable laser.

The capacity and density of such a memory are ultimately determined by how sharply the absorption bands are synthesized and how well the relative positions in the spectra are controlled. The success of this concept depends also on the availability of tunable laser sources with linewidths compatible with the material requirements.

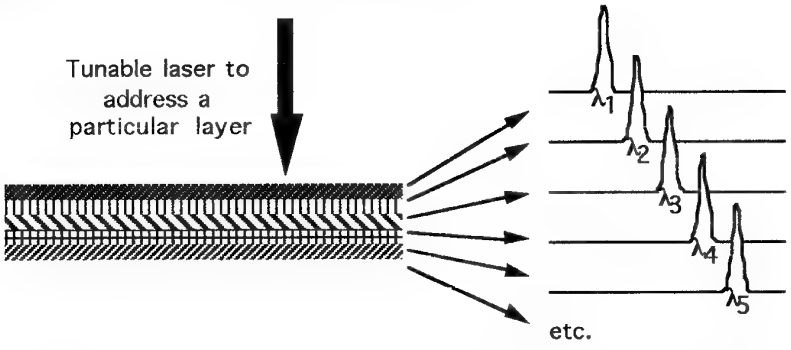


Figure 12: The hypothetical spectra of the various films used a multi-wavelength memory with many layers.

5.7 Two-photon 3-D memory concept

Two photon three dimensional memory[54,55,56] is a data storage technique that requires the simultaneous absorption of two photons in order to store information in the material. A simple diagram showing how this method can create a high capacity memory is given in Figure 13.

The physical process at the heart of the three dimensional memory (3DM) is a molecular change caused by a two photon optical absorption, as shown in Figure 14 for a spiropyran molecule. A molecule in the ground (unwritten) state is excited to a higher energy state by the simultaneous absorption of two distinct photons, one red (1064 nm) and one green (532 nm). The energy required to reach the excited state is greater than either photon alone can provide, but when two photons interact simultaneously they are absorbed, resulting in a bond dissociation. The molecular geometry (structure) is changed into a new, written, molecule with an entirely different absorption spectrum. The intensity of the infrared beam can be high, because a two photon absorption of the infrared beams has insufficient energy to write the material. However, a relatively low intensity of green light, in the presence of the intense infrared beam, can write the material.

The written bit can be read by illumination with two photons of a different energy than those absorbed by the unwritten molecule (i.e., at 1064 nm, as shown in Figure 14). The written molecule absorbs the two 1064 nm photons and fluoresces in the red, at around 700 nm. Using two photon absorption to write and read the memory makes it possible to identify a single bit anywhere within a three-dimensional volume by simply intersecting two optical beams at that point. The capacity of such a bit-oriented volume memory is fundamentally limited only by the memory volume divided by the optical spatial resolution, leading to an upper bound in data storage density as high as 10^{12} bits/cm^3, compared to 10^8 bits/cm^2 for planar storage media.

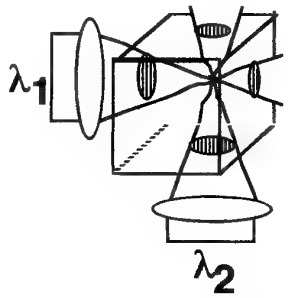


Figure 13: The schematic diagram for using the two photon effect to write throughout a volume of material.

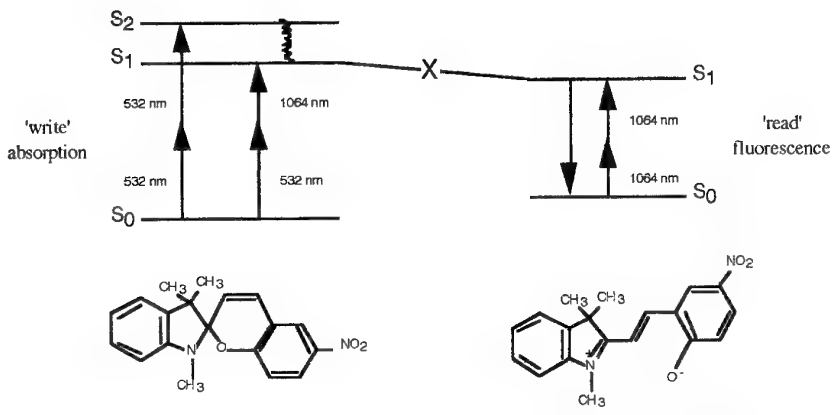


Figure 14: Two photon absorption (left) and readout (right) processes for a spiropyran molecule.

Most of the past research for the secondary storage materials was based on bond dissociation of spirobenzopyran (SP) molecules held in a polymer host. A more detailed discussion of these results can be found in a recent publication[57]. The initial challenges in materials development were to find materials that (a) could be written and read using the two photon effect and (b) exhibited the long term stability of the written form critical for secondary storage. Several promising two photon materials have been identified and incorporated into a polymer host material (PMMA) which can be shaped into any form appropriate for the optical system. The absorption and fluorescence spectra of SP1 in a polymer host are shown in Figure 15. The recording energy required for such materials is about 10 pJ/μm^3, and the fluorescence efficiency is on the order of 1%. The storage lifetime depends on temperature and the readout intensity. The dark storage lifetime is in the months at 3°C, and years at 77°K.

The written form of SP1 is a polar molecule, with a positive and negative charge on the ends of the open ring. The storage lifetime can be extended indefinitely by chemically binding the charged ends to another polar molecule or by anchoring them to the polymer matrix. We have demonstrated nearly infinite stability of the written form by using HCl to bridge the written (open) forms. These materials were tested and bits were recorded in the memory cube. The absorption and fluorescence spectra of the bridged SP1 molecule are shown in Figure 16. Note that the fluorescence peak is very well separated from the absorption peaks. This is important for practical application of the material, since if the fluorescent output were absorbed by the written material, crosstalk would occur between a recalled output and other planes of stored data. The bridged materials will be suitable for a write-once, read many times memory and can be used for secondary and archival storage systems.

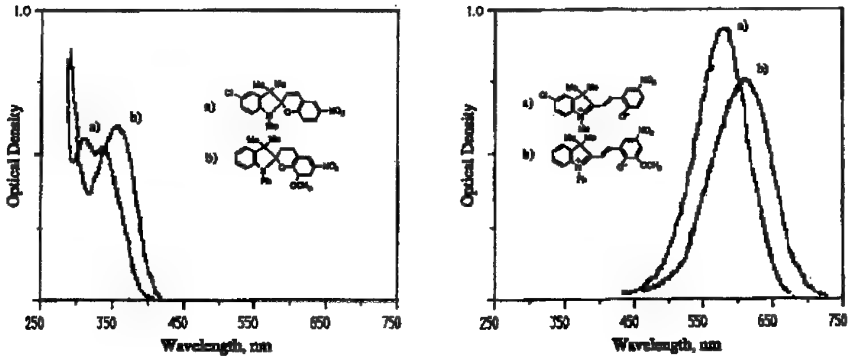


Figure 15: Spectral response for the absorption of the unwritten (I) and written (II) forms of SP1, as well as the fluorescence spectrum of the written form (III).

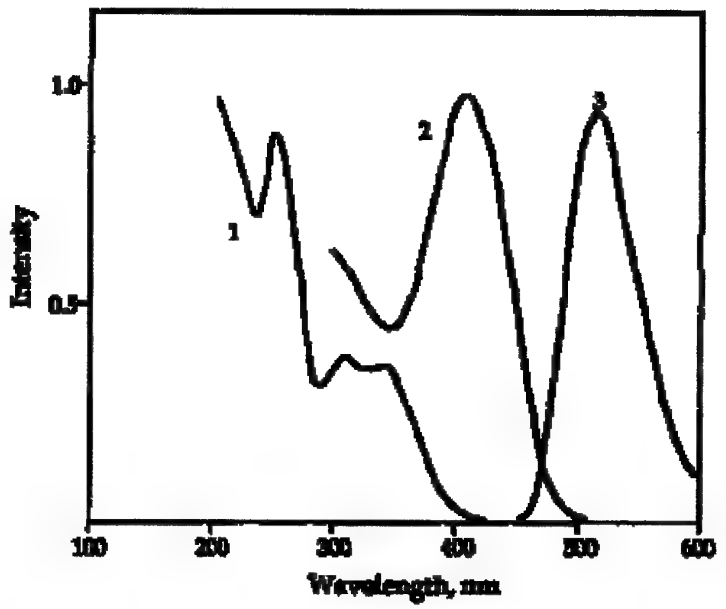


Figure 16: Spectral response for the bridged SP1 material, showing the absorption of the unwritten (I) and written (II) forms of SP1, as well as the fluorescence spectrum of the written form (III).

A second approach to using two photon absorption for 3-D memories that modifies the refractive index change in the material has also been demonstrated.[58] This method writes the information into the material using one sharply focused beam to cause two photon absorption near the focal point. Differential interference contrast microscopy is used to convert the phase modulation into an intensity modulation and read the data. Both reading and writing of the material are limited in the amount of parallel data that can be written using this second approach.

6. SYSTEM DESIGN: TWO PHOTON 3-D MEMORY

In the previous section, we have introduced several approaches to 3-D memories. In this section we focus on one approach, that of two-photon 3-D memory, and investigate potential system architecture and the required devices that will be suitable for its implementation. A two-photon 3-D memory can be addressed by simply intersecting two focused beams, but this only allows bit sequential readout. For parallel address, the data is organized into images or binary bit-planes. These bit planes can be stored within the memory using two different data addressing methods: orthogonal and counterpropagating pulse collision addressing.

6.1. Orthogonal Addressing system

Orthogonal addressing (beam intersection) is based on the interaction of a 2-D optical image containing the data to be stored entering the memory cube from one face with a second enable 1-D field that defines the bit plane at orthogonal incidence, as shown in Figure 17. This scheme can in principle use cw lasers as light power sources, although two photon processes are much more efficient at the higher intensities characteristic of pulsed laser sources. Figure 18 shows in more detail the components necessary to build an actual memory device. The principle components

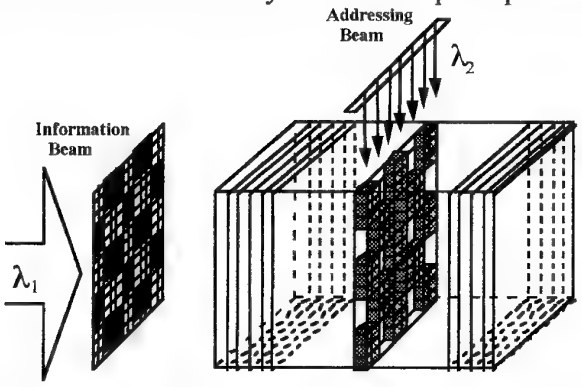


Figure 17: Image storage in two photon 3-D memory materials using orthogonal beam addressing.

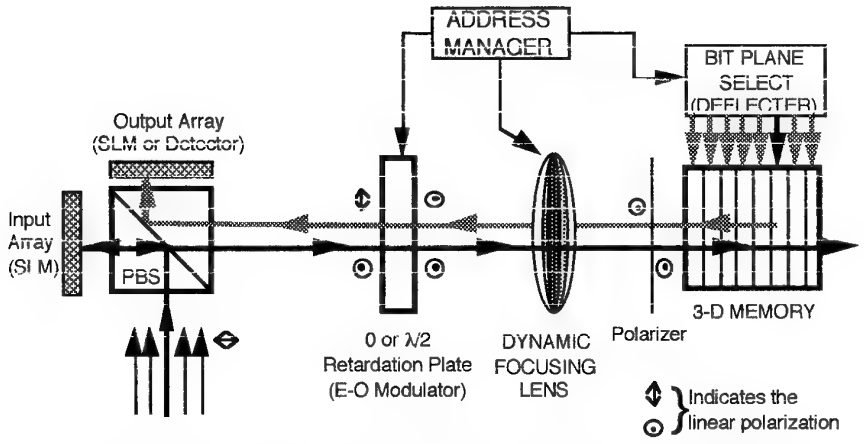


Figure 18: System diagram of orthogonal 3D memory system using orthogonal beam addressing.

required are a spatial light modulator to compose the data and a dynamic focusing lens to image the data onto the desired memory plane, a beam deflector to direct the enable beam onto the storage planes, and an output detector array.

It should be possible to store a one dimensional line of data with diffraction limited resolution and storage density. For two dimensional arrays, however, orthogonal addressing can not simultaneously provide high capacity and high data rates because the minimum width of the enable beam is restricted by Gaussian beam propagation of the cylindrical 'enable' beam. A low F/number lens will illuminate a small area with a narrow line, or a larger F/number lens can illuminate a large area with a deeper beam. The deeper the 'enable' beam, the higher the parallelism possible, but the larger the volume which must be dedicated to each bit. For example, to maintain a data storage density of 1 Gbit/cm^3, the parallelism must be limited to approximately 128x128.

The experimental system for recording an image using two photon absorption is shown in Figure 19. A flashlamp pumped Nd:YAG laser producing 30 ps pulses was frequency doubled to produce output at 1064 and 532 nm. The energy per pulse was approximately 1 mJ in both the green and the infrared pulses. However, the green output was used to illuminate a relatively large area input plane (about 100 mm^2) while the infrared output was focused to a line approximately 8 mm by 100 μm so that its intensity was much higher. The secondary memory material used was an unpolished cube of SP1 in PMMA polymer. The input image was transmitted through the cube and imaged onto the CCD camera for recording the output.

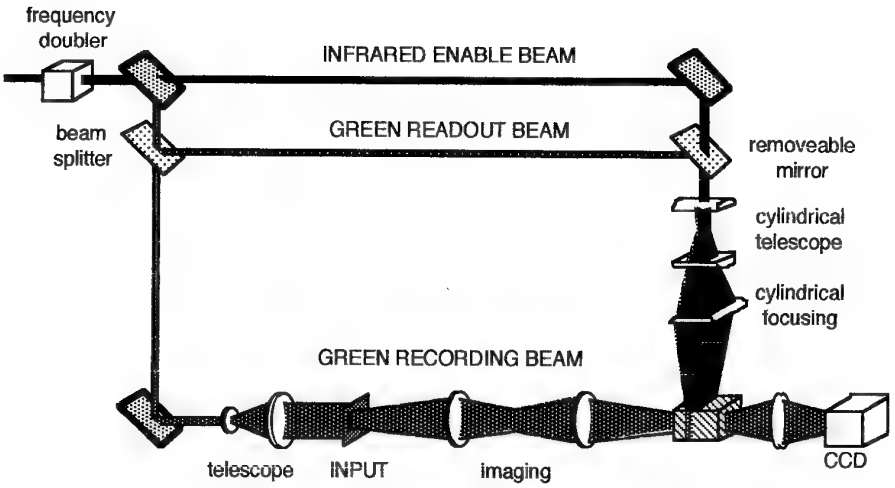


Figure 19: Experimental image recording set up.

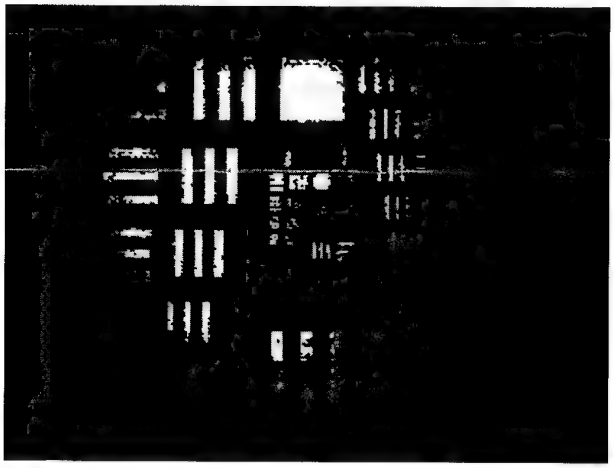


Figure 20: Transmitted (left) and recalled (right) image stored by two photon absorption.

Figure 20a shows the transmitted image. The recording plane was selected by intersecting the infrared line with the green image beam within the cube. The images could in principle be recalled using two photon absorption of the infrared beam. However, to increase output intensity a small fraction of the green beam was diverted along the infrared path to allow recall by single photon absorption. The recalled image is shown in Figure 20b. The resolution seemed to be comparable to that of the transmitted image, although there was a significant amount of background intensity which reduced the contrast ratio.

We have chosen a dual approach in constructing the dynamic focusing lens, combining a continuously variable liquid crystal lens[59] with a

discrete stepping holographic lens.[60] The holographic lens can provide quick changes between a small number of widely separate focal planes, while the liquid crystal lens provides slower access to a continuum of more closely spaced planes. The concept of the holographic dynamic focusing lens (HDFL) is shown in Figure 21. Basically, a number of lens functions are computed as computer generated holograms. Each is multiplied with a unique random phase pattern. Then the holograms are summed together and the result is fabricated as a surface relief pattern to yield a multiplexed phase-only hologram. The individual holograms are accessed by imaging one of the encoding phase patterns onto the hologram plane. The justification for this approach is that a relatively simple electro-optic spatial light modulator (for example, a PLZT wafer patterned with surface electrodes) can access the high resolution holographic lens in a very short time.

An HDFL set-up was experimentally demonstrated. The phase code addressing for these tests was provided using etched glass phase plates, rather than a electro-optic SLM. Using a single point input an output spot size of 8.5 μm microns was measured with a signal to background intensity ratio of greater than 100:1. The resolution and uniformity were limited by the holographic spot array generator used rather than by the HDFL.

6.2 Pulse collision addressing system

Pulse collision addressing was devised to provide higher data transfer rates using more parallel data channels. An ultra-short pulse tunable infrared Ti:Sapphire laser provides the illumination. Infrared pulses 100 fs (1 fs = 10^{-15} s) long are routinely generated by such lasers, and pulses as short as 17 fs have been demonstrated.[61] A fraction of the laser output is frequency doubled into the green and used to illuminate a spatial light modulator (SLM) displaying the data plane for storage. This bit-plane is sent into one face of the two photon material. The rest of the laser output is sent into the memory from the opposite face. The two pulses intersect within the material, and the information is stored. The thickness of the stored plane is determined by the pulse widths. Since the speed of light in a material of index n is 0.3/n μm/fs, the intersection of the two 100 fs pulses in a material with an index of 1.5 is approximately 20 μm thick. The position of the intersection is determined by the path length of the two pulses. Each plane is addressed by applying a corresponding physical delay to one of the beams. Memory readout is also accomplished by pulse collision, using wavelengths corresponding to the absorption spectra of the written bits. An additional advantage to short pulse lasers is the increase in

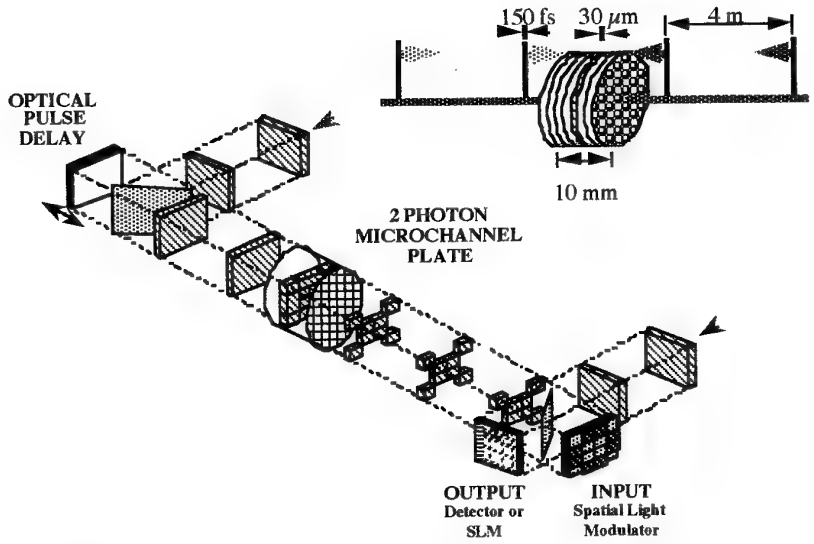


Figure 23: Two photon storage using ultrashort laser pulse collision addressing.

two photon absorption probability which occurs at the high peak intensities characteristic of such lasers. The geometry for pulse collision recording is shown in Figure 23.

In addition to the increase in data density, a major advantage of pulse collision is that the dynamic focusing lens can be eliminated without loss of resolution by incorporating the storage material into a waveguide array. Microchannel arrays used for image amplifiers are made by etching the central cores from a glass optical fiber bundle, leaving an array of hollow channels. Theliquid storage material can be drawn into the channels then allowed to harden. The resulting waveguide array will transmit an image from the front surface to the back without blurring the image.

Pulse collision addressing within the microchannel or bulk material is accomplished by applying a physical path delay to one of the recording or readout beams. One possible method of doing this uses an acousto-optic modulator to reflect the beam from stepped mirror. The short pulse beam has a significant color spectral width ($\Delta\omega$ is about 1% for a 100 fs beam) which increases as the pulse width decreases. The acousto-optic modulator changes this color spectrum into an angular spectrum which could smear the beam. However, using a two-pass geometry corrects for this dispersion exactly. The result can be a cascadeable random access pulse delay with up to 1 MHz access speeds.

For the two photon memory using pulse collision addressing, the data retrieval (access) time is limited by the speed at which the position of the colliding pulses can be controlled. We propose to develop an optical pulse

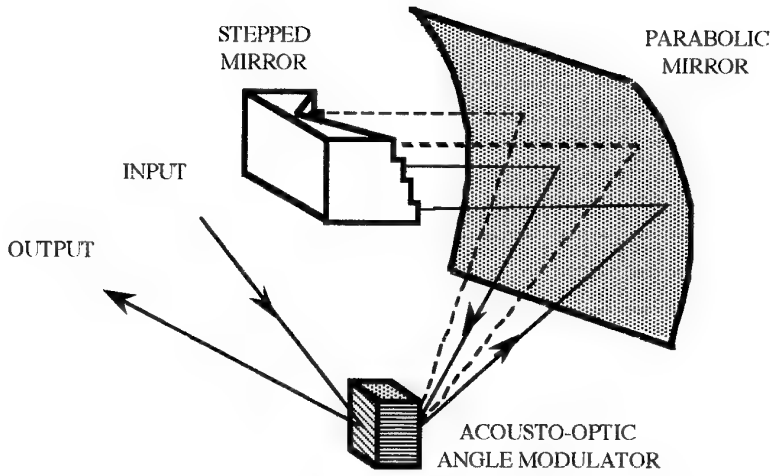


Figure 24: Achromatic short-pulse optical pulse delay using acousto-optic modulation.

delay (OPD) that can provide an access time down to 0.1μs. The characteristics of an ultra-short pulse laser output (high peak intensity and a 5-10 nm wavelength bandwidth) render most approaches impractical. Figure 24 shows a proposed OPD device which can provide the necessary performance. The input pulses diffract from a small aperture acousto-optic deflector. The diffracted light reflects from a parabolic mirror onto a V-grooved stepped mirror constructed with the required delay settings. The reflected light is re-diffracted by the A-O cell, automatically correcting for diffractive pulse dispersion. Currently available commercial A-O cells are available with 5 MHz access time and 112 resolvable spots. Using two such units, cascaded together, an OPD could yield over 1024 distinct intersection planes within the memory material. For high speed operation near the 100 MHz range, standard A-O cells have fewer resolvable spots (<10). Therefore several units must be cascaded to increase the total number of addressable memory planes to reach the desired number of planes. Ultimately, the entire device could be constructed with integrated optics using surface acoustic wave (SAW) modulators to increase operation speed, reduce the size, and enhance reliability.

7. CONCLUSIONS

It is expected that over the next few years for some specific applications areas such as associative memories a fast transition from sequentially addressed optical disk systems to parallel accessed disk systems will occur to satisfy their high data rate and medium capacity requirements. However, 3-D optical memories may have a profound impact on the present storage hierarchy. A considerable amount of

research and development over the next decade will be needed however, to take 3-D memories from concept to technology. We strongly believe that such an effort is well justified by the applications in fast, high capacity storage.

At this early stage, when major improvements in 3-D optical material characteristics are expected, it is premature to assess the far reaching technological implications of many of the 3-D memory concepts. However, the fast progress made over the past few years in PSHB, multi-wavelength and two-photon storage raises our hope that by the turn of the millennium Terabyte capacity optical memories could become available with access times of less than a millisecond and data transfer rates exceeding Terabits/sec.

This is a shorter version of a chapter on optical storage that can be found in SPIE CR47-06 "SPIE Critical Technology Review Series"

8. REFERENCES

1 A. E. Bell, "Critical issues in high-density magnetic and optical data storage," Laser Focus **19**, 61-66, (1983).

2 K. Sato, K. Fujita, M. Miyazawa, M. Shirai, K. Kobayashi, M. Ishihara, T. Nakao, "A system-integrated ULSI chip containing 11 4Mb RAMs, 6 64kb SRAMs and an 18k gate array," 1992 IEEE International Solid-State Circuits Conference, 52-53, (1991).

3 H. Suguira, E. Morita, S. Nagasawa, "F6631 solid state disk: high speed virtual disk unit," Fujitsu Scientific and Technical Journal 26, 296-305 (1991).

4 S. A. Przybyski, *Cache and Memory Hierarchy Design*, Morgan Kaufmann Publishers, Inc., San Mato, CA, 1990.

5 E. G. Ramberg, "Holographic Information Storage," RCA Review **33**, 5-53 (1972).

6 K. Kubota, Y. Ono, M. Kondo, S. Sugama, N. Nishida, and M. Sakaguchi, "Holographic disk with a high data transfer rate," Applied Optics **19**, 944-051 (1980).

7 D. Psaltis, M. Neifeld, A. Yamamura, and S. Kobayashi, "Optical memory disk in information processing," Applied Optics **29**, 2038-2057 (1990).

8 J. Rilum and A. Tanguay, "Utilization of optical memory disk for optical information processing" in *OSA annual Meeting*, Vol. **11** of 1988 OSA Technical Digest Series (Optical Society of America, Washington, D.C., 1988), paper M15.

9 P. Marchand, A. Krishnamoorthy, K. Urquhart, P. Ambs, S. Esener, and S. H. Lee, "Motionless-head parallel readout optical disk system," Applied Optics **32**, 190-203 (1993).

10 W. P. Altman, G. M. Claffie, and M. L. Levene, "Optical storage for high performance applications in the late 1980s and beyond," RCA Engineer **31**, 46-55 (1986).

11 5 1/4" Write-Once, Read-Many Optical Disk Specification Sheet, Daicel Chemical Industries, Torrance, CA, 1991.

12 A. Krishnamoorthy, P. Marchand, G. Yayla, and S. Esener, "Opto-electronic associative memory using parallel readout optical storage," UCSD Internal Rep. (University of California, San Diego, La Jolla, CA, 1991) and submitted for publication in IEEE Transactions on Neural Networks.

13 W. J. Smith, *Modern Optical Engineering*, Chapter 6, (McGraw-Hill, Inc. 1990).

14 R. G. Zech, "Volume hologram optical memories," Optics and Photonics News, 16-24, (Aug. 1992).

15 R. Collier, C. B. Burckhardt, and L. Lin, *Optical Holography*, (Academic Press, 1971).

16 L. Solymar and D. J. Cooke, *Volume Holography and Volume Gratings*, (Academic Press, 1981).

17 P. M. Hariharan, *Optical Holography*, (Cambridge University Press, 1984).

18 F. T. S. Yu, S. Wu, A. Mayers, S. Rajan, and D. A. Gregory, "Color holographic storage in LiNbO3," Opt. Comm. **81**, 348-352 (1991).

19 P. J. van Heerden, "Theory of optical information storage in solids," Appl. Opt. **2**, 393-400 (1963).

20 I. J. Cox and C. R. J. Sheppard, "Information capacity and resolution in an optical system," JOSA A, **3**, 1152-1158 (1986).

21 K. Bløtekjaar, "Limitations on holographic storage capacity of photochromic and photorefractive media," Applied Optics **18**, 57-647 (1979).

22 F. Mok, D. Psaltis, and G. Burr, "Spatially- and angle- multiplexed holographic random access memory," SPIE Proc. **1173**, 334-345 (1992).

23 D. Psaltis, D. Brady, and K. Wagner, "Adaptive optical networks using photorefractive crystals," Applied Optics **27**, 1752-1759 (1988).

24 A. Marrakchi, W. M. Hubbard, S. F. Habiby, and J. S. Patel, "Dynamic holographic interconnects with analog weights in photorefractive crystals," Opt. Eng. **29**, 215-224 (1990).

25 Y. Taketomi, J. Ford, H. Sasaki, J. Ma, Y. Fainman, and S. H. Lee, "Incremental recording for photorefractive hologram multiplexing," Opt. Lett. **16**, 1774-1776, 1991.

26 P. Gunter and J. P. Huignard, Eds., *Photorefractive materials and their applications I and II*, (Springer–Verlag, 1988 and 1989).

27 G. C. Valley and M. B. Klein, "Optimal properties of photorefractive

materials for optical data processing," Opt. Eng. **22**, 704-711 (1983).

28 G. C. Valley, M. B. Klein, R. A. Mullen, D. Rytz and B. Wechsler, "Photorefractive materials," Ann. Rev. Mater. Sci. **18**, 165-188 (1988).

29 H. Kurz, "Photorefractive recording dynamics and multiple storage of volume holograms in photorefractive $LiNbO_3$," Optica Acta **24**, 463-473 (1977).

30 P. Gunter and A. Krumins, "High-sensitivity read-write volume holographic storage in reduced $KNbO_3$ crystals," Appl. Phys. **23**, 199-207 (1980).

31 D. Von der Linde, A. M. Glass and K. F. Rodgers, "High–sensitivity recording in KTN by two–photon absorption," Appl. Phys. Lett. **26**, 22-24 (1975).

32 J. E. Ford, Y. Fainman, and S. H. Lee: "Enhanced photorefractive performance from 45°-cut $BaTiO_3$," Applied Optics **28**, 4808-4815 (1989).

33 J. E. Ford, J. Ma, Y. Fainman, S. H. Lee, et al, "Multiplex holography in SBN:60 with applied field," JOSA A **9**, 1183-1192 (1992).

34 J. W. Burgess, R. J. Hurditch, C. J. Kirby, and G. E. Scrivener, "Holographic storage and photoconductivity in PLZT ceramic materials," Applied Optics **15**, 1550-1557 (1976).

35 F. Mok and H. M. Stoll, "Holographic inner product processor for pattern recognition," Proc. SPIE **1701**, 312 (1992).

36 R. V. Johnson and A. R. Tanguay, Jr., "Fundamental physical limitations of the photorefractive grating recording sensitivity," Chapter 3, *Optical Processing and Computing*, (Academic Press, 1989).

37 A. J. Smirl, K. Bohnert, G. C. Valley, R, A. Mullen, and T. F. Boggess, "Formation, decay, and erasure of photorefractive gratings written in $BaTiO_3$ by picosecond pulses," JOSA B **6**, 606-615 (1989).

38 D. L. Staebler, W. J. Burke, W. Phillips, and J. J. Amodei, "Multiple storage and erasure of fixed holograms in Fe-doped $LiNbO_3$," Appl. Phys. Lett. 26, 182-184 (1975).

39 B. J. Chang, "Dichromated gelatin holograms and their applications," Opt. Eng. **19**, 642-648 (1980).

40 W. K. Smothers, B. M. Monroe, A. M. Weber, and D. E. Keys, "Photopolymers for holography," *Practical Holography IV, SPIE OE/Lase Conference Proceedings* **1212-03**, (1990).

41 A. M. Weber, W. K. Smothers, T. J. Trout, and D. J. Mickish, "Hologram recording in Du Pont's new photopolymer materials," *Practical Holography IV, SPIE OE/Lase Conference Proceedings* **1212-04** (1990).

42 R. R Birge, "Photophysics and molecular electronic applications of the rhodopsins," Annu. Rev. Phys. Chem. **41,** 683-733 (1990)

43 R.R. Birge, Biochim. Biophys. Acta **1016**, 293-327 (1990)

44 F. V Bunkin, N.N. Vsevlodov, A.B. Druzhko, B.I. Mitsner, A.M. Prokhorov, V.V. Savranskii, N.W. Tkachenko and T.B. Shechenko, Sov. Tech. Phys. Lett. **7**:630-631 (1981).

45 R.B. Gross, K. C. Izgi and R.R. Birge, SPIE **1662**, Image Storage and Retrieval Systems, pp 186 (1992).

46 W.E. Moerner, ed., *Persistent Spectral Hole-Burning : Science and Applications* (Springer-Verlag, Berlin 1988).

47 C. De Caro, A. Renn, U. P. Wild, "Spectral hole-burning - applications to optical image storage," Berichte Der Bunsen Gesellschaft Fur Physikalische Chemie **93**, 1395-1398 (1989).

48 C. Brauchle, "Spectral hole burning at room temperature and with a single molecule - 2 new perspectives," Angewandte Chemie-International Edition in English **31**,426-429(1992).□

49 S. Arnold, C. T. Liu, W. B. Whitten, J. M. Ramsey, "Room-temperature microparticle-based persistent hole-burning spectroscopy," JOSA B-Optical Physics **9**, 819-824 (1992).

50 C. De Caro, A. Renn, U. P. Wild, "hole burning, stark effect, and data storage 2. Holographic recording and detection of spectral holes," Applied Optics **30**, 2890-2989 (1991).

51 B. Kohler, S. Bernet, A. Renn, U. P. Wild, "Holographic optical data storage of 2000 images by photochemical hole burning," Persistent Spectral Hole-Burning: Science and Applications, OSA Conference Monterey, California **16**, 46-49 (1991).

52 T. W. Mossberg, "Time-domain frequency-selective optical data storage," Optics Letters **7**, 77-79 (1982).

53 International Symposium on Future Electron Devices - Bioelectronic and Molecular Electronic Devices - [FED BED/MED SYMPOSIUM], November 20-21, 1985, Tokyo.

54 D. M. Parthenopoulos and P. M. Rentzepis, "Three-dimensional optical storage memory," Science **245,** 843-845 (1989).

55 S. Hunter, F. Kiamelev, S. Esener, D. A. Parthenopoulos, and P. M. Rentzepis, "Potentials of 2-photon based 3-D optical memories for high performance computing," Applied Optics **29**, 2058-2066 (1990).

56 A. S. Dvornikov, S. Esener, and P. Rentzepis, "3-Dimensional Optical Storage Memory by means of two-photon interaction," Optical Computing, S. H. Lee and J. Jahns, Ed., Wiley Academic Press (1993).

57 P. M Rentzepis, A. S. Dvornikov, "Three dimensional optical memory by means of two-photon processes," presented at OE-LASE 1993, SPIE Proceedings **1153**, Los Angeles, CA (1993).

58 J. H. Strickler, W. W. Webb, "3-dimensional optical data storage in refractive media by 2-photon point excitation," Optics Letters **16**, 1780-1782 (1991).

59 P. Brinkley, S. Kowel, and C. Chu, "Liquid crystal adaptive lens: beam translation and field meshing," Applied Optics **27**, 4578 (1988).

60 S. Hunter and S. Esener, "Dynamic focusing lens for volume memory applications," SPIE Proceedings **1773**-15 (1992).

61 C. P. Huang, M. T. Asaki, S. Backus, M. Murnane, H. Kapteyn, and H. Nathel, "17-fs pulses from a self-mode-locked Ti:sapphire laser," Optics Letters **17**, 1289-1291 (1992).

The Emerging Field of Artificial Neural Networks and Their Optoelectronic Implementations

Aharon J. Agranat
Department of Applied Physics
The Hebrew University of Jerusalem
Jerusalem 91904, ISRAEL.

Summary

Artificial neural networks based on models which were developed in the context of brain research, are becoming a significant data processing tool. Neural computing algorithms are robust, parallel, can be trained from examples, and perform associative memory recall. Special purpose hardware is essential for implementing these algorithms effectively. Combined opto-electronic implementations of these models seem to be the preferred embodiments. Two approaches for implementing artificial neural networks by a combined opto-electronic systems will be described: The optical disk based artificial neural networks, and the Electroholographic artificial neural networks.

Introduction

Over the past decade the field of neural computation has become the focus of intense interdisciplinary research in computer science, neurobiology, physics and engineering. Originally neural network (NN) theories were developed in the context of brain research, as a theoretical tool conceived for understanding the principles governing the operation of the nervous system. However, the special properties of these models brought forward the possibility of harnessing them to the solution of various artificial intelligence tasks. These properties are the following:

1. The models can perform associative memory recall.
2. The models are robust and fault tolerant. Namely, part of the an NN can be

Paper presented at the AGARD SPP Lecture Series on "Optical Processing and Computing" held in Paris, France from 12-13 October 1995; Rome, Italy from 16-17 October 1995 and Ankara, Turkey from 19-20 October 1995 and published in LS-199.

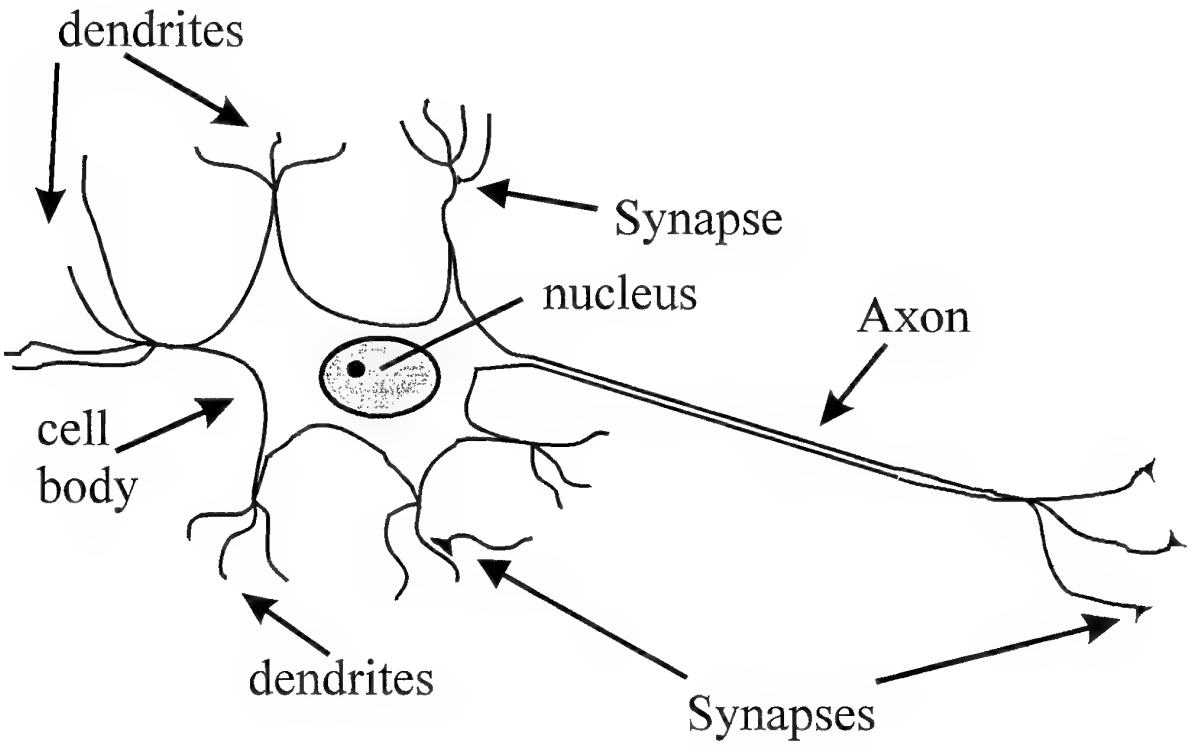


Figure 1: Schematic drawing of a typical neuron.

damaged without noticeable effect on its functioning capability.

3. The models are flexible and self adjusting. A NN can tune itself to perform a certain task by learning from given examples. For example, a network can train itself to perform a certain classification task by scanning a small part of the input-output space, and applying a 'learning' procedure to tune itself. The network can now classify the entire input-output space, although most of it was not scanned in the training process (as required in conventional expert systems when the decision tree is built).
4. NN can deal with information which is noisy and probabilistic.
5. NN are highly parallel.

Thus the field of artificial neural network (ANN), emerged as a by product of brain research, but is evolving independently as a generic computing paradigm. It should therefore be emphasized that the usefulness of NN theories to data processing and computing is independent of their relevance to neuroscience.

Parallel to the theoretical research of neural computing, and the exploration of its potential applications, much effort was devoted to the invention and development of special purpose hardware of ANN. Early on it was realized that simulations of neural computing paradigms on digital hardware are limited. Digital hardware is tailored to meet the needs of digital computing which is a coarse grain high accuracy task, whereas neural computing is a fine grain low accuracy task. Special purpose hardware was therefore developed to meet the needs of neural computing. Naturally most of the effort was devoted to microelectronic implementations of ANN. It was however found out that the incorporation of optics is essential for the implementation of very large scale ANN.

In the next section a brief description of some NN models will be given, illustrating their characteristic properties as data processing tools. Following this section, two examples of optoelectronic ANN will be presented. The first system is the optical disk based neural network system. The second system is the electro-holographic ANN which is based on a new concept in optical computing: Electroholography (EH). EH provides a direct interface between electronic circuits and volume holographic devices.

Essentials of ANN

Since neural computing is inspired by neuroscience, some knowledge of the basic neurobiological nomenclature is required. The basic building block of the nervous system is the nerve cell or the **neuron**. There

are many different types of neurons, however, NN theories are mainly concerned with a stereotypical neuron as described schematically in Figure 1. The cell body (or **soma**), of the neuron receives stimuli primarily through the **dendrites**, which are tree like networks of nerve fiber connected to the soma. The neuron transmits signals through the **axon** which is a 'transmission line' extending from the soma. The axon is connected to dendrites of other cells, or directly to their somas by interface elements - the **synapses**. Signals transmission between neurons is a complex chemical process, occurring at the respective synapse, which induces a change in the electrical potential inside the cell body of the receiving neuron. Once this potential exceeds a certain threshold, the neuron fires a pulse (called **action potential**) along its axon. There are approximately 10^{11} nerve cells in the brain, each connected to 10^3-10^4 other neurons. As mentioned above there are many different types of neurons, and many of the fine details which distinguish them are omitted in the stereotypical picture presented above. A concise review of neuroscience can be found in Ref. (1).

Inspired by this schematic picture, ANN are ensembles of processing elements called **neurons**, interconnected by interface units called **synapses**. The state space of the neuron is either discrete or continuous. Discrete neurons are in most cases binary neurons that are either active (in state '1'), or non active (in state '0'). The state space of continuous neurons is some continuous segment (cf. [0, 1]). The dynamics (or the update process) of a neuron is a process in which the weighted sum of the signals it receives is computed, and a new state is assigned to it according to this sum.

Consider a network in which the i-th neuron is in a state designated by V_i. This neuron is updated to a new state according to the following procedure:

$$\tilde{V}_i = \Phi(h_i, p_i) \qquad [1a]$$

$$h_i = \sum_j^{M_i} W_{ij} V_j \qquad [1b]$$

where h_i is the post synaptic input, Wij is the interaction strength of the synapse that interfaces neuron j to neuron i, and there are M_i neurons which are interfaced to the i-th neuron. Φ is the decision process by which the neurons are updated. If the neuron is binary Φ is a threshold function such as

$$\Phi(h) = \begin{cases} 1 & \text{if } h \geq 0 \\ 0 & \text{if } h < 0 \end{cases} \qquad [2]$$

Φ for graded response neurons is some nonlinear function, normally the sigmoid function

$$g(h) = \frac{1}{1+e^{-\beta h}} \qquad [3]$$

Φ can also be a non deterministic process , in which case

$$Prob\{ V_i = 1 \} = \Phi(h_i) \qquad [4]$$

namely, Φ is the probability that the new state of the i-th neuron be '1' .

Several different architectures of ANN exist, by which neurons are grouped together to form networks. Most architectures can be divided into two classes: feed forward (FF) networks, and recurrent networks.

FF networks (sometimes referred to as **multi layers perceptrons**), are networks in which the neurons are arranged in layers, and the output from one layer is fed into the input of the next layer (Figure 2a). The layers are normally referred to as the input layer, the first hidden layer, the second hidden layer etc., and the output layer. The direction of the signal from the input layer through the hidden layer and finally out of the output layer is well defined. Thus, a FF network is a nonlinear transformation of the form

$$V_O = \Phi(V_I) \qquad [5a]$$

$$\Phi: \quad \Re^n \rightarrow \Re^m \qquad [5b]$$

where V_I is the input vector, V_O is the output vector, and n and m are the dimensions of the input and output layers (vectors) respectively.

As an example consider Figure 2b. The XOR Boolean logic function is implemented by a network composed of an input and output layers and one hidden layer.

FF networks are effective for three different functions: (a) FF networks can implement Boolean logic functions (a capability which has a theoretical significance but not necessarily a practical value); (b) FF networks are useful for various pattern recognition applications. In particular, as a classification tool for partitioning the patterns space into categories using 'supervised learning' methods; and (c) FF networks can be efficiently used for implementing nonlinear transformations of functional approximation problems.

One of the reasons for the usefulness of FF networks as a classification tool, is the fact that they can be trained by scanning a small fraction of their input domain. The training algorithms are classified into two categories: supervised learning and unsupervised learning. Supervised learning

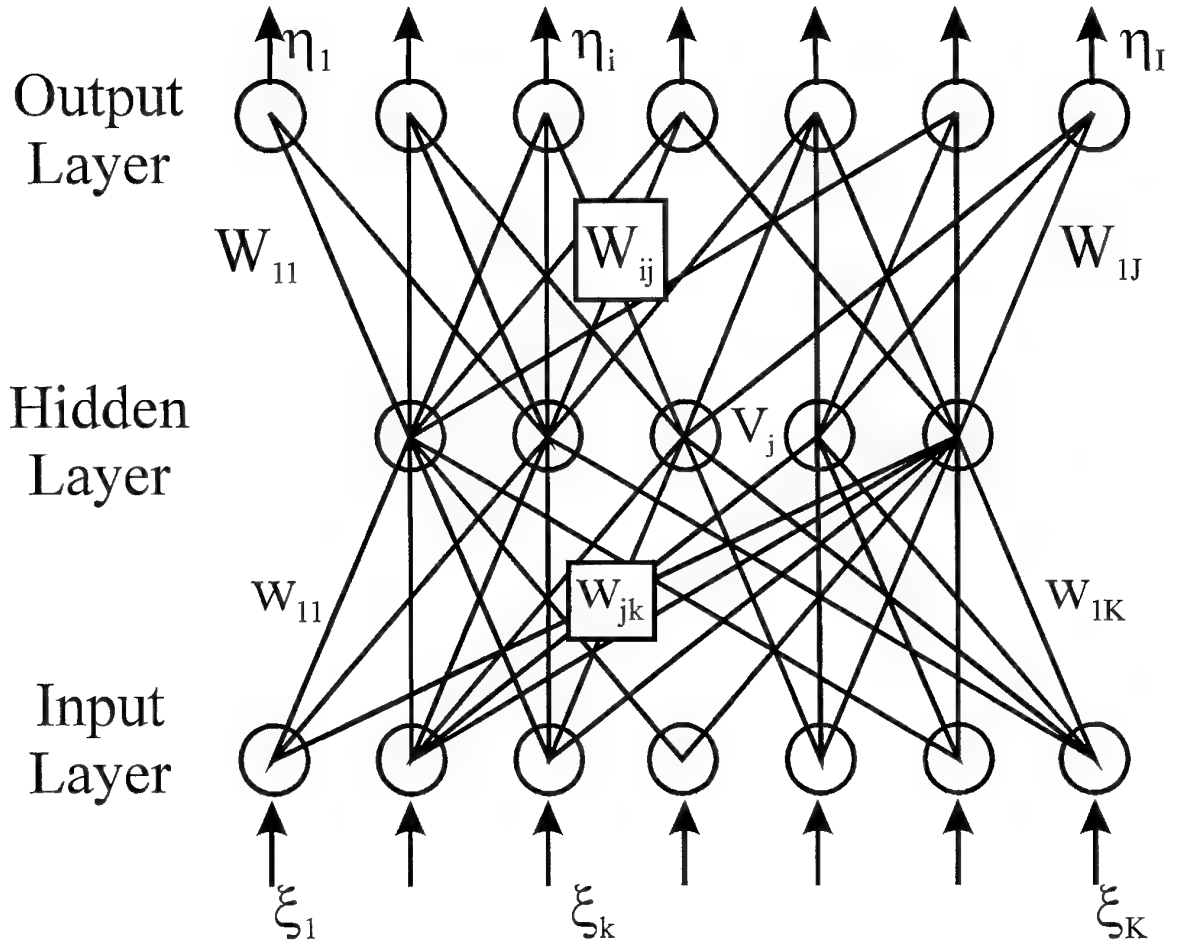


Figure 2a: A feed forward neural network with one hidden layer.

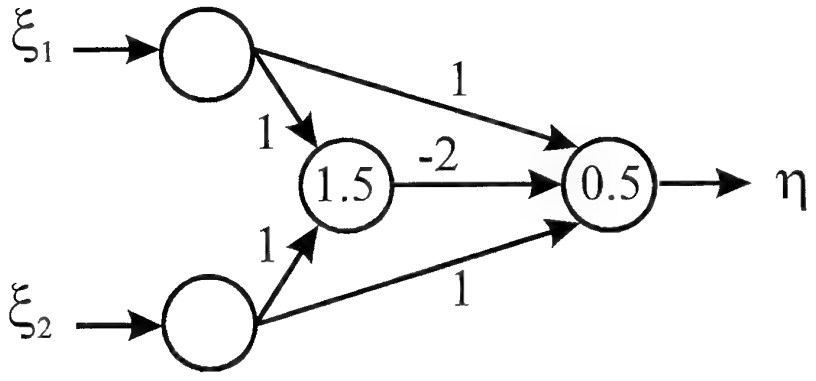


Figure 2b: A feed forward network implementing the XOR function

is a training process in which training set containing a set of inputs with their respective 'correct' outputs is known. Thus the network can measure the level of its performance and adapt itself accordingly to perform the task at hand.

Consider Figure 3 in which a schematic description of supervised learning is presented. The network is fed with an input from a training set $\{\xi^\mu , \zeta^\mu\}$ where ζ^μ, $\mu=1,...p$ is the correct output for the input ξ^μ. The produced output vector η^μ, is then used together with the correct output ζ^μ (extracted from the training set), to derive improved values for the network synapses. The process is repeated until the network learns the training set, namely for each input vector ξ^μ that is fed into the network, the required output vector ξ^μ is produced.

An example of the supervised learning process is the 'Back Propagation' algorithm, which is also the most widely used NN training algorithm. Consider a two layers network as described in Figure 2a. The input vector is fed into the hidden layer by the synaptic weights (w_{kj}), and the output from the hidden layer are fed into the output layer by the synaptic weights (W_{ij}). The training procedure is a gradient descent process in which the cost function given by

$$E[W] = \frac{1}{2}\sum_{\mu,i}(\zeta_i^\mu - \eta_i^\mu)^2 \qquad [6]$$

is minimized in the synaptic weights space. The minimization is accomplished by an iterative process in which at each iteration the weights are updated according to

$$w_{jk} := w_{jk} + \Delta w_{jk} \qquad [7a]$$

$$W_{ij} := W_{ij} + \Delta W_{ij} \qquad [7b]$$

It was found that the correction terms for the synaptic weights of the output layer ΔW_{ij}, are given by

$$\Delta W_{ij} = \eta \sum_\mu \delta_i^\mu V_j^\mu \qquad [8a]$$

$$\delta_i = g'(h_i^\mu)(\zeta_i^\mu - \eta_i^\mu) \qquad [8b]$$

and similarly for synaptic weights of the hidden layer

$$\Delta w_{jk} = \eta \sum_\mu \delta_j^\mu \xi_k^\mu \qquad [9a]$$

$$\delta_j = g'(h_j^\mu)\sum_i W_{ij}\delta_i^\mu \qquad [9b]$$

Consider δ_j used for computing the correction terms Δw_{jk} of the hidden layer interconnects. It can be easily seen that δ_j is produced by letting the error δ_i flow in the

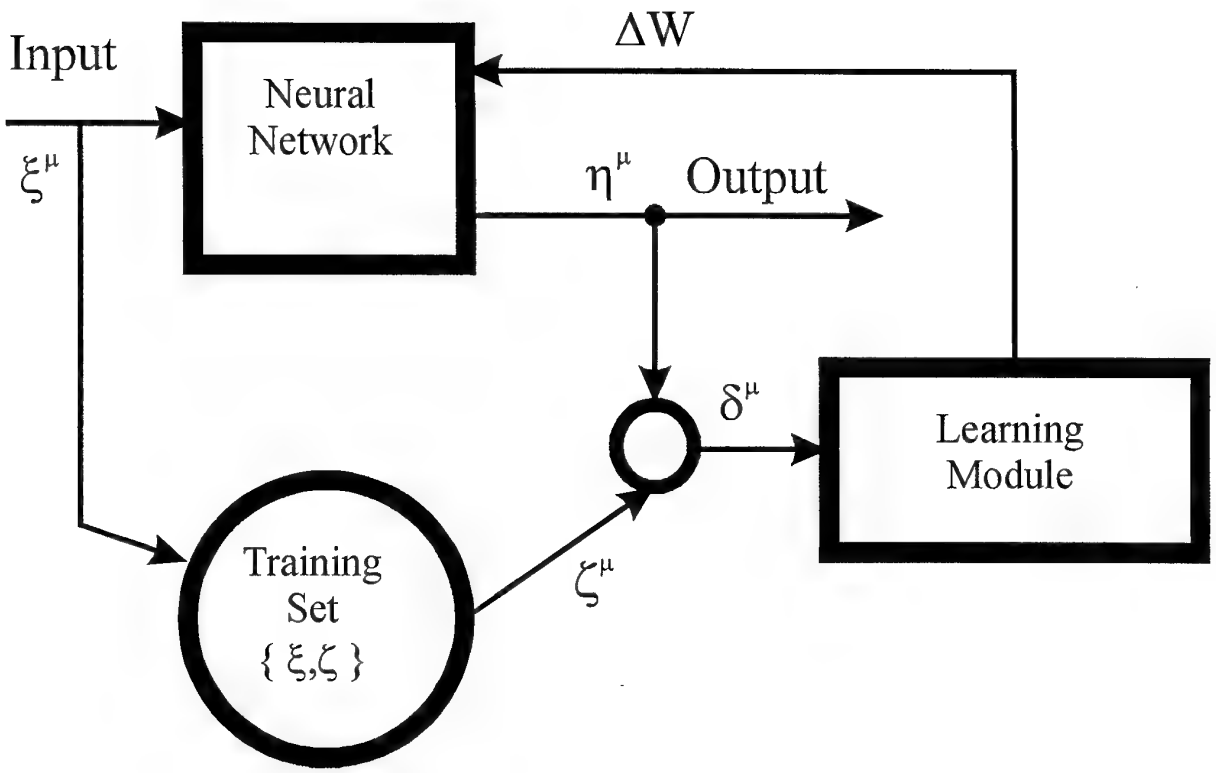


Figure 3:A block diagram of a 'Supervised Learning' training process.

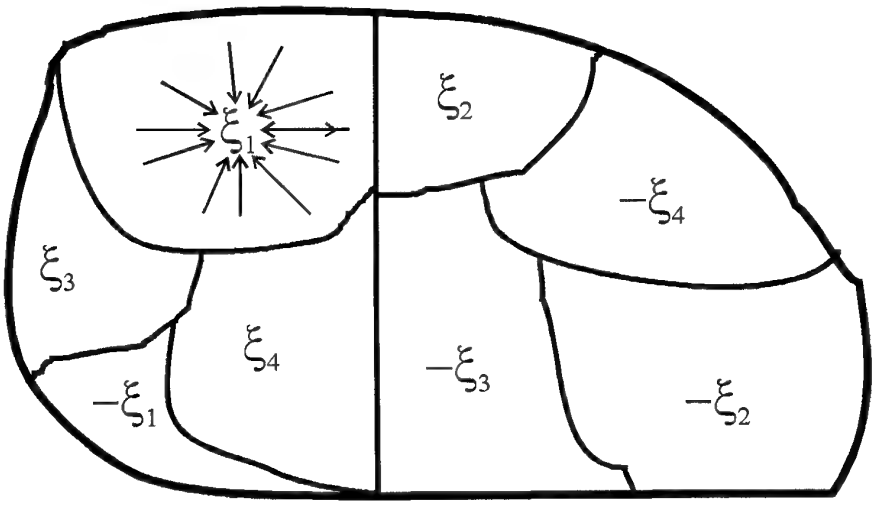


Figure 4: A schematic description of the Basin of Attraction of a Hopfield network with four stored patterns.

network in the reverse direction, hence the term Error Back Propagation (EBP).
Advanced versions of EBP are the most widely used NN training algorithms in many pattern recognition applications.

Recurrent networks are dynamic networks in which each neuron can in principle be connected to all other neurons. The node equations of these networks are described by differential equations. Recurrent networks evolve in time and can either converge to a particular stable state, travel randomly through the state space, or converge into a subset of the state space.
Recurrent networks are useful for a wide variety of computational tasks such as system modeling, predictions of time evolution of system behavior, sequence recognition, and trajectory following.

As an example consider the application of the Hopfield network to the solution of the associative memory problem. The Hopfield model describe a NN of fully interconnected binary neurons. For mathematical convenience the states of the neurons will be designated by "1" (firing), and "-1" (non firing). The dynamics of the network can now be written

$$\tilde{V}_i = \mathrm{sgn}\,(h_i - U_i) \qquad [10]$$

where h_i is defined in [1], U_i is the threshold level, and sgn(x) is the sign function defined as

$$\mathrm{sgn}\,(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \qquad [11]$$

The associative memory problem can be formulated as follows:
A set of p patterns which are N-dimensional binary vectors of the form

$$\xi^{\mu} = (\xi^{\mu}_1, \ldots, \xi^{\mu}_N) \qquad \mu = 1,\ldots p \qquad [12]$$

are stored in the memory. To which of these p prestored patterns does a given pattern ζ resembles the most?
(Where resemblance is defined as the pattern to which the Hamming distance, is minimal. Namely , the pattern with which the number of identical bits is maximal).
Obviously this problem can be solved serially by computing the Hamming distance from the given pattern ζ , to each of the p prestored patterns which is given by

$$\sum_{j=1}^{N} [\xi^{\mu}_j (1-\zeta_j) + (1-\xi^{\mu}_j)\,\zeta_j] \qquad [13]$$

and evaluating the pattern ξ^{ν} for which the Hamming distance is minimal. The Hopfield model solves this problem in a different way. In the learning stage the patterns are stored by assigning values to the

synaptic interaction matrix according to

$$W_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_i^{\mu}\xi_j^{\mu} \qquad [14]$$

It can be easily seen that if a network defined according to [10], the patterns ξ^{μ} are stable states. Namely, if the network is in one of the stated ξ^{μ} it will remain there. Moreover, it can be seen that if the system is set to a particular set ζ , close to one of the prestored patterns ξ^{μ} , it will (under certain assumptions) converge to the said pattern. This situation is illustrated schematically in Figure 4 where it can be seen that the state space of the network is divided into basins of attraction of the various stable states.

In the Hopfield network not all the stable states are the prestored patterns. For example, if ξ^{μ} is a prestored stable state, then - ξ^{μ} is also stable. Thus the performance of such networks as content addressable or associative memories is not perfect. In particualr, it depends on the ratio p/N.

When this ratio exceeds a certain level, spurious states which are stable will be formed. This brings about the usefulness of non deterministic, or stochastic, neurons. Networks of such neurons can free themselves from some spurious states and improve their performance.

Finally it should be emphasized that the models and algorithms presented in this section constitute a small fraction of the wide field of NN models. An exhaustive general review of NN models and their applications can be found in ref. (2), and a concise review of supervised learning algorithm can be found in ref. (3).

Hardware Implementations of ANNs:

At the present time, the flexibility and versatility of operation of digital computers make them the preferred research tool for simulating neural models. However, a typical NN is a massively interconnected network of low accuracy processors, whereas a digital Von Neumann computer consists of one or a few sparsely interconnected sophisticated high accuracy processors. Therefore, using the latter for implementing the former, results in a tradeoff between size and speed of operation of the implemented network.

This inherent inefficiency in performance of ANN implemented on digital computers, prevents neural computing from becoming a viable computing technology. A massive R&D effort was therefore launched for creating special purpose neural hardware.

Naturally, since microelectronics is the dominating signal processing technology, most of this effort was devoted to the

construction of ANN on silicon based VLSI chips.

Part of this development effort is summarized in Table 1 which is extracted from Reference (4). The state of the art has not improved significantly since reference (4) was published in 1991. (Although some of the projects which were at the development stage in then, have been either completed or discarded).

Table 1: VLSI Neural Network Implementations Existing in 1991

	SWUPS (*)	SW Accuracy	Number of Synapses	Number of Neurons	Technology	Synapse Area (µm2)
AT&T	80B	1b x 16b	8K - 32K	256	9 mm CMOS	5100
Adaptive Solutions	1.6 B	1-16b x 1-16b	128 K - 2M	64	.8 µm CMOS multi field dye	1400
CALTECH (Agranat)	0.5 B	5b x 5b	65536	256	2 µm CCD	560
Intel (Holler)	2B	6b x 6b	10240	64	1 µm cmos EEPROM	2009

(*) SWUPS: Synaptics Weights Updates (Multiply Accumulate Operations) per Second

Note that some of the development groups belong to the leading microelectronic giants, and the most advanced technologies have been used, however, very small networks have been developed. The reason for this is the fact that the capacity of electronic networks is inherently limited by the planar design of silicon VLSI chips. The basic building block of ANN is the synapse. Therefore, the dimension of a fully interconnected ANN is proportional to the square root of the area of the given silicon 'real estate'.

A typical synapse performs a 'multiply accumulate' operation therefore its physical dimensions are similar to those of a multiplier. A very small and simple synapse is 10x10 μm^2. Thus a fully interconnected network implemented on 5x5 cm^2 of silicon 'real estate' will contain 5000 neurons at the most.

This limitation brought forward the possibility of incorporating optics in hardware implementations of ANNs. The key advantages of optics are its abilities to provide the required massive interconnection network. This is achieved by combining the parallel operation of optical free space interconnections with the gigantic capacity of optical memories.

In the next sections two approaches for constructing ANNs are presented, in which electronic circuits are interconnected by optical free space interconnects. The first approach is based on an existing technology: optical memory disk, to store the synaptic weights. The second approach is more futuristic, and is based on a generic concept: electroholography which enables the direct interface between electronic circuits and holographic interconnects.

Example 1: The Optical Disk Based Neural System

The optical disk based ANN is a hybridization of an electronic processor with an optical memory, combining the advantages of electronics in computing, with the superiority of the optical disk in data storage capacity and retrieval rate.

A close analysis of the general dynamics of NN reveals that the main computational burden is the computation of the post synaptic input (h_i in Equ. (1lb.)). Computing the post synaptic input requires a vector - matrix multiplication.

It differs however, from the similar operation required for discrete transforms computations (e.g. the Discrete Fourier Transform), which is a high accuracy multiplication of a vector by a matrix of fixed values. For neural computing low accuracy is sufficient, while the matrix of the synaptic weights is often very large, and changes between successive operations.

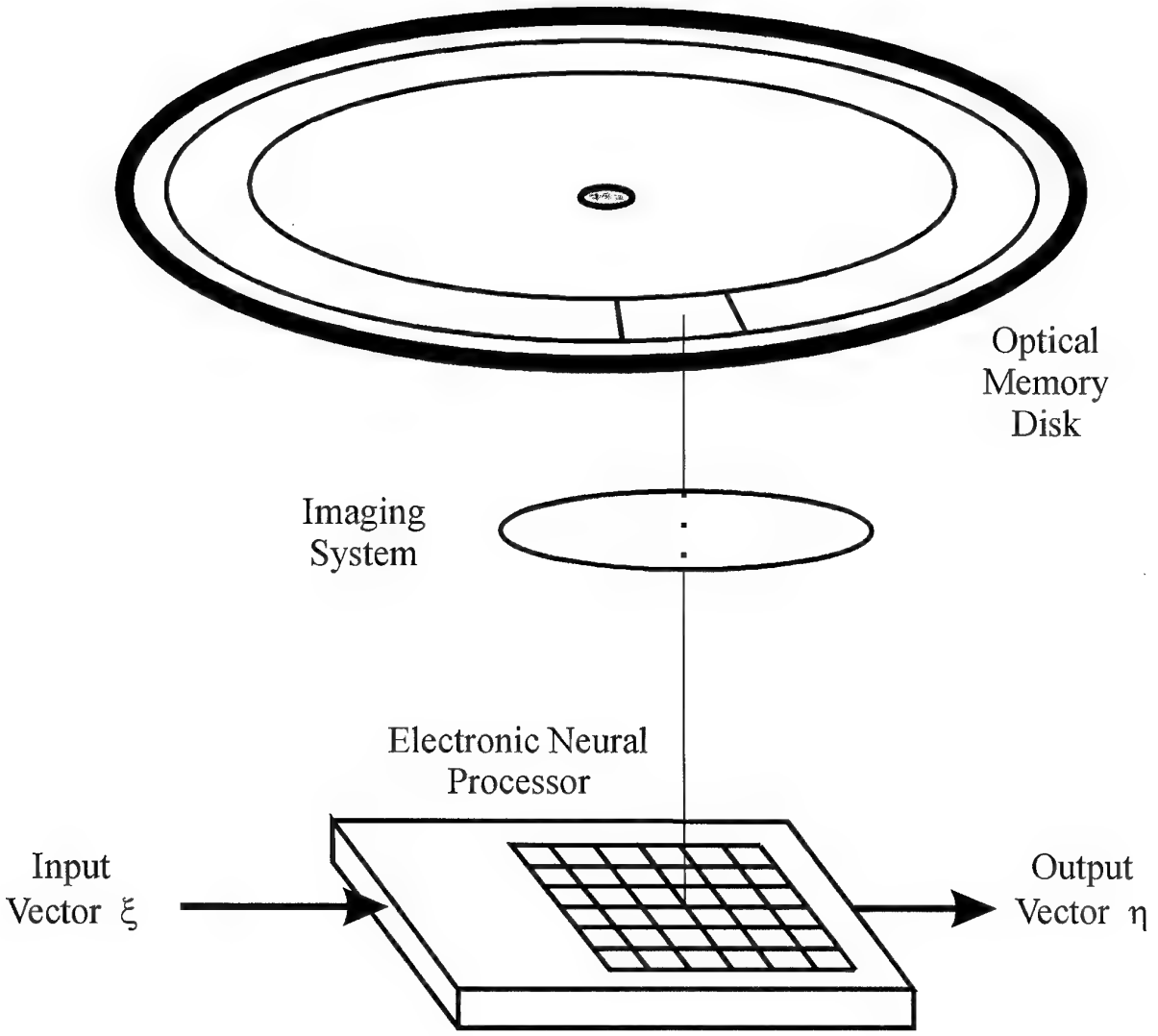


Figure 5: A schematic description of the optical disk based ANN.

Performing vector - matrix multiplications under this constraint is very inefficient when purely electronic hardware is used. Optics is incorporated to overcome this inefficiency. Consider figure 5 in which a schematic layout of the optical disk based is presented The synaptic interaction matrix (or part of it), is encoded in the form of an image. The light emitted by each pixel of such an image is proportional to the synaptic weight of the respective term in the synaptic matrix. (e.g. the light emitted by the (i,j) pixel is proportional to Wij). Thus the synaptic matrix is fed into the electronic neural processor (NP) in the form of a 2D distribution of an optical signal. A detectors array (DA) acting as the receiving unit of the NP, transforms this signal into a 2D electrical signal. The NP now multiplies the newly loaded matrix by the input vector which is fed to it electronically.

The idea to load an electronic NP from an optical memory was first proposed by Agranat et al. (5), (6). A series of electronic NPs were developed using charge coupled devices ((5), (7), and (8)), polysilicon detectors (9), and NP photodiodes implemented in a CMOS process (9). Most noteworthy among the NPs is the CID - NP which can in principle achieve a computing rate of 10^{12} multiply accumulate operations per second, (8) (11).

Originally it was proposed to use a spatial light modulator (SLM) to load the matrix into the electronic NP, and to use an optical memory to store the synaptic interaction matrix. As such, this architecture is very limited, since the bottle neck is simply shifted from the link between the SLM and the DA, to the link between the electronic memory and the SLM which remains serial. It was then proposed by Psaltis et al. (12) to prestore a set of synaptic matrixes as images on an optical disk, and to image them onto the NP using the appropriate synchronization mechanism. This approach optimizes the combination of the electronic NP with the optical memory since the memory and SLM are integrated into one device - the disk.

A small prototype of this system was built at Caltech by Psaltis and co-workers. (12). The system contain a NP based on photodiodes implementd in CMOS technology, and a Sony CD prototype system. Based on their preliminary experiments it is estimated that this approach will lead to data transfer rate of 35 Gb/sec. (12).

Example 2: The Electroholographic ANN

In the previous example the optics and electronics are integrated at the system level. An electronic processor is loaded from an optical memory system. While it remains advantageous to exploit the advantages of

combining electronic processors by optical free space interconnects, it is desirable to perform the integration at the devices level. The generic concept of electroholgraphy (EH) provides exactly that capability. EH enables interconnecting electronic neurons by minute volume holograms, using the voltage controlled photorefractive effect in paraelectric crystals (13).

It is clear, however, that in order to simulate different functional units in the brain (cf. one orientation column in the primary visual cortex (V1)) at least 10,000 neurons are needed. The future need for compact implementations of very large scale ANN, is the underlying motivation behind the EH ANN

The photorefractive effect enables the recording of optical information on crystals, by changing the local index of refraction in response to light energy it absorbs. The information is recorded in the form of phase holograms that can be retrieved by applying the reconstructing (reading) light beam at the appropriate wavelength and/ angle. In the paraelectric phase one can control the efficiency of the effect by applying an external electric field to the crystals during the recording stage, and through the reading phases.

In general, the diffraction efficiency is proportional to the local photoinduced changes in the birefringence ($\delta(\Delta n)$). At the paraelectric phase the electrooptically induced birefringence depends quadratically on the electric field and is given by:

[15]

$$\Delta n(x) = \tfrac{1}{2} n_0^3 g \varepsilon_0^2 \varepsilon^2 (E_0 + E_{sc}(x))^2$$

where $\Delta n(x)$ is the induced change in the index of refraction, n_0 is the refractive index, g is the quadratic electrooptic coefficient, and it is assumed that the polarization is in the linear region $P = \varepsilon_0 \varepsilon E$, where E is the electric field and ε is the dielectric constant. Let $E=E_0+Esc(x)$, where E_0 is the externally applied field and Esc(x) is the photoinduced spacecharge field.

The change which contributes constructively to the diffraction is given by:

[16]

$$\delta(\Delta n(x)) = n_0^3 g \varepsilon_0^2 \varepsilon^2 E_0 E_{sc}(x).$$

Thus it can be seen that the information carrying spacecharge field is transformed into a local change in refractive index only in the presence of an external electric field. Therefore the use of the quadratic electrooptic effect enables an analog control of the storage and reconstrucion of information. Recently a new crystal: potassium lithium tantalate nibate (KLTN) was developed. KLTN doped with copper and vanadium was found to be particularly

suitable to be used as the nedium for EH devices:

1. In KLTN the work point can be set to be at room temperature. Slightly above the phase transition temperature Tc, the dielctric constant ε is very large ($\varepsilon = 10^4$-10^5), so that moderate electric fields will induce a large photorefractive effect. Therefore it is desirable to set the work point slightly above Tc. KLTN crystals were grown in which Tc is slightly below room temperature, while maintaining high optical quality.
2. In KLTN in the proximity of the phase transition it is possible to create fixed photorefractive gratings which are not erased by the reading light.
3. The photorefractive sensitivity of KLTN is approximately 10^{-4} cm^3/J , an order of magnitude superior to $LiNbO_3$.

Thus the voltage controlled photorefractive effect in KLTN provides us with a natural tool for controling light beams by electronic circuits. The electroholographic (EH) neural network exploits this capability.

Each neuron (Figure 6), is an independent electronic circuit performing a decision function based on its input (for example a nonlinear electronic amplifier), its output (Vi for the i-th neuron), is applied to a minute photorefractive crystal. Prior to the operation of the network, one hologram that contains all the synaptic connections for that neuron ({Wij} j=1, 2, ... for the i-th neuron), is stored in the crystal. Since the diffraction efficiency is proportional to the external electric field, the intensity of the output image, which is diffracted by the crystal, will be proportional to the product of the neuron's activity (Vi) by the synaptic strengths {Wij} . The image is detected by a an array of linear detectors.

The complete network contains a two dimensional array of these "pixel neurons" and a detectors array, as shown in Figure 7. All of the holograms are recorded on all of the crystals with only one reference beam, allowing a parallel reconstruction of all the images on the detectors. Each detector performs a linear summation of the subjected light intensity from all the respective neurons i.e.: $\sum_j W_{ij} V_j$. This matrix-vector multiplication is performed within one time constant of the detector. It is important to note that the technology for integrating KLTN pixels on a silicon wafer was developed by Texas Instruments for their uncooled pyroelectric detectors arrays.

Any neural architecture can be implemented using the electroholographic concept, if we connect the input to each neuron to the respective detector, we will describe a recurrent neural network. But the detector can be connected to another layer in the network, providing us with a feed-forward

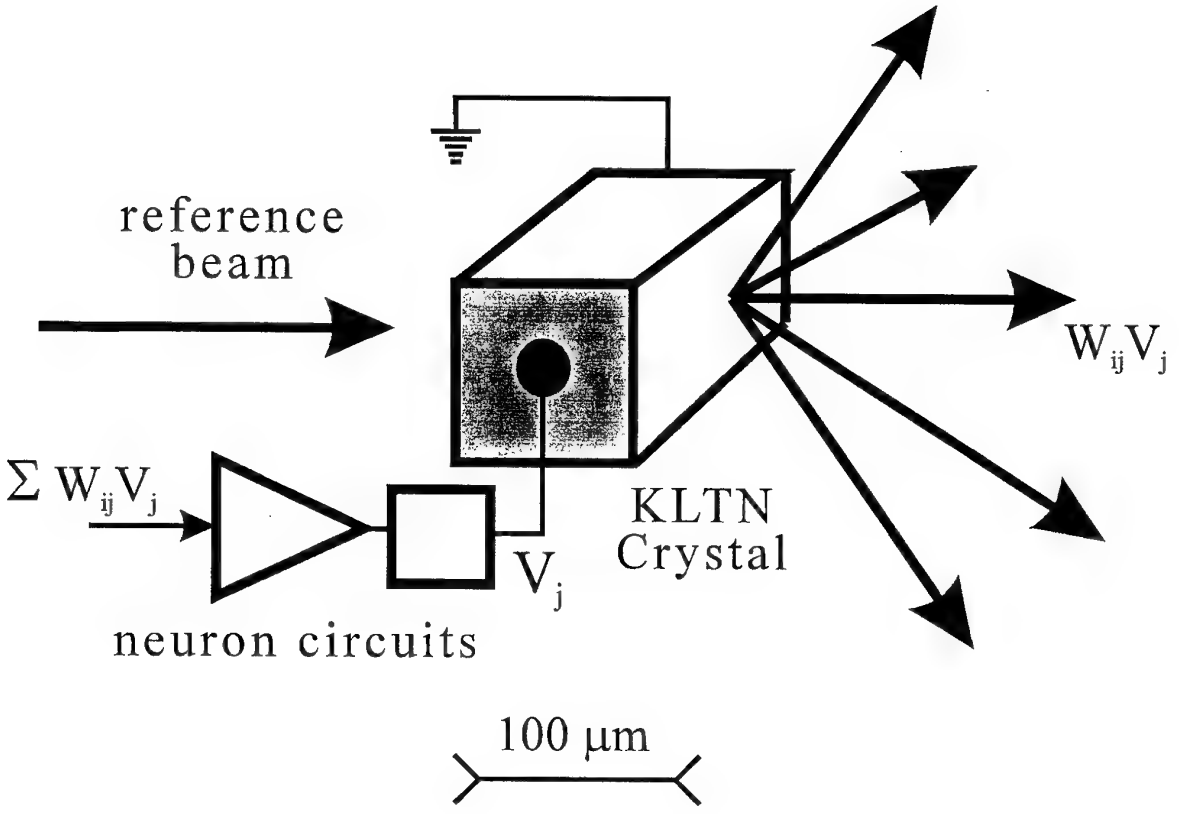


Figure 6: Schematic Description of One Neuron in an Electroholographic Neural Network.

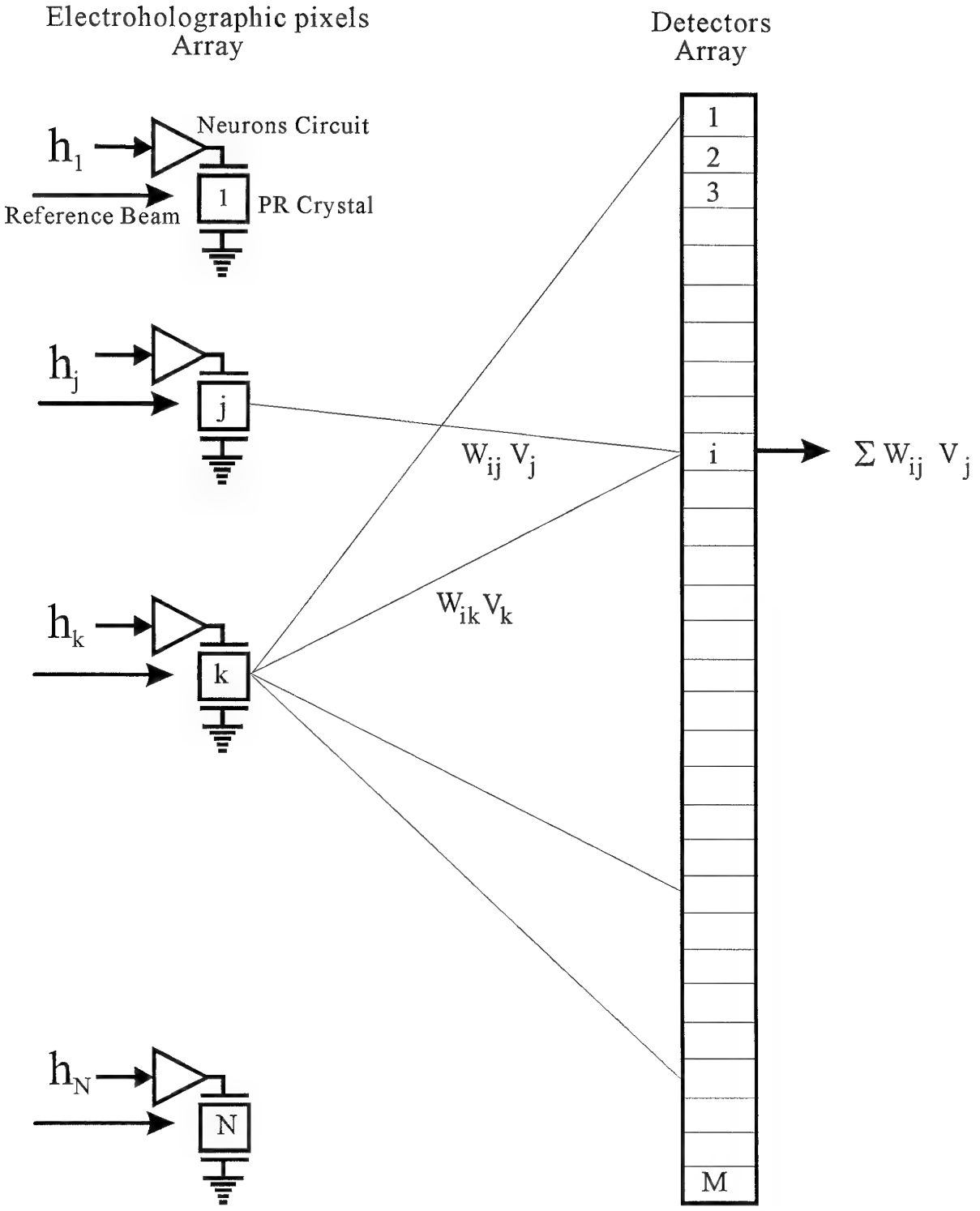


Figure 7: Basic architecture of electroholographic neural network.

architecture. The trade-off between electronic connections and optical connections allows one to connect small (about 1000 neurons) electronic functional units with each other into a larger network (100,000 neurons) where the long-range connections are optical. The connection update rate of such a network can be easily estimated: For a ANN in which the EH pixels page contains 500x500 pixels, each connected to 1000 synapses, assuming that the input is fed into the EH pixels page at video rate (10Mhz), the expected update rate is $2.5 \bullet 10^{15}$ connections updates per second !

A prototype of the EH ANN is currently being built at the Hebrew University of Jerusalem in Israel.

Conclusions:

ANNs are potentially a powerful information processing tool. Special purpose hardware implementation of ANN are required in order to realize this potential. Implementations which combine electronic data processing with optical memories and communication systems, seem to be the most advantageous approach for achieving this goal.

References:

(1).T. J. Sejnowski, and P. S. Churchland, "The Computational Brain", MIT Press (1992).

(2).J. Hertz, A. Krogh, and R. G. Palmer, "Introduction to the theory of neural Computation", Addison Wesley (1991).

(3). D. R. Hush, and B. G. Horne, "Progress in supervised Neural Networks", IEEE Signal Processing mag. **10**, 8 (1993).

(4). M. A. Holler, "VLSI implementations of learning and memory systems: A review", in ***Adv. in Neural Information Processing Systems*** **III,** (R. P. Lippman et al. Eds.), Morgan Kaufmann Publishers (1991).

(5). A. J. Agranat, C. F. Neugebauer, and A. Yariv ,'Spatial Light Modulators As Parallel Memories for Optoelectronic Neural Networks', PROC. SPIE 1150 ,(1989),

(6). A. J. Agranat, C. F. Neugebauer and A. Yariv, 'Parallel Optoelectronic Neural Network Processors' , **US Patent 5,008,833 (1991).**

(7). A. J. Agranat and Y. Yariv, 'Semiparallel Microelectronic Implementation of Neural Network Models Using CCD Technology',

Electronics Lett. **23** , 580 (1987).

(8). A. J. Agranat, C. F. Neugebauer, and A. Yariv ,'A Parallel Optoelectronic Implementation of Neural Network Models Using CID Technology',

Appl. Opt. **27** ,4354 (1988).

(9) K. Kornfeld et al. in the ***Proc. International Conf. on Neural Networks***, Washington DC, (January 1990)

(10). C. F. Neugebauer, A. J. Agranat, and A. Yariv, 'Optically Configured Phototransistor Neural Networks' , International Joint Conf. on Neural Networks 1990, Washington D.C., (1990).

(11).C.F.Neugenbauer and A. Yariv, "The CID neural network chip ", in ***Adv. in Neural Information Processing Systems*** **IV,** (J. E. Moody et al. Eds.), Morgan Kaufmann Publishers (1992).

and:

A. J. Agranat, C. F. Neugebauer, and A. Yariv 'A Charge Domain Bit Serial Vector Matrix Multiplier for Real-Time Signal Processing',**U.S. Patent 5,258,934 (1993).**

(12). A. A. Yamamura, M. A. NeifeldS. Kobayashi, and D. Psaltis, "Optical disk based artificial neural system", Optical Computing and Processing 1, 3 (1991).

(13). A. J. Agranat, V. Leyva, and A. Yariv, 'Voltage Controlled Photorefractive Effect in Paraelectric $KTa_{1-x}Nb_xO_3$: Cu,V',

Opt. Lett. **14** , 1017 (1989).

(14). A .J. Agranat, 'Electroholographic Artificial Neural Networks', Physika **A 200**, 608-612 (1993).

(15). A. J. Agranat, R. Hofmeister and A. Yariv, 'Characterization of a New Photorefractive Material: $K_{1-y}Li_yTa_{1-x}Nb_xO_3$'

Optics Letters **17**, 713 (1992) .

(16). R. Hofmeister, S. Yagi, A. Yariv, and A. J. Agranat, 'Growth and Characterization of KLTN:Cu,V Photorefractive Crystals',

J. Cryst. Growth **131**, pp 486-494 (1993).

Ultra-Fast Nonlinearities in Semiconductor Optical Amplifiers for Applications in All-Optical Networks

Kerry J. Vahala, Jianhui Zhou*, Namkyoo Park**
California Institute of Technology
Pasadena, California 91125
USA

Mike Newkirk*** and Barry Miller
AT&T Bell Laboratories
Holmdel, New Jersey
USA

* Current address: AT&T Bell Laboratories, Holmdel, NJ
** Current address: AT&T Bell Laboratories, Murray Hill, NJ
*** Current address: Ortel Corporation, Alhambra, CA

Summary

Ultra fast intraband occupancy relaxation in semiconductor gain media has recently been shown to provide a wide band nonlinearity which is several orders of magnitude larger than the Kerr nonlinearity in silica fiber. We address recent work directed towards applications of this nonlinearity to the wavelength conversion function in all optical networks; specifically, carrier wavelength spectral translation by four-wave mixing. In addition to reviewing the current performance of these devices including conversion efficiency, signal to noise and a simple system demonstration, we will discuss the underlying physics of the ultra-fast four-wave mixing mechanism and its application to TeraHertz spectroscopy of intraband scattering. An overview of wavelength conversion in the context of all optical networks is provided and competing techniques to four-wave mixing wavelength conversion are also discussed.

Paper presented at the AGARD SPP Lecture Series on "Optical Processing and Computing" held in Paris, France from 12-13 October 1995; Rome, Italy from 16-17 October 1995 and Ankara, Turkey from 19-20 October 1995 and published in LS-199.

I. Introduction

1.1 Overview and Definitions

A wavelength converter is a device which translates information on an optical carrier at one wavelength to a new desired wavelength. As explained below, it will be a critical function in future fiber networks [1,2] and for this reason there are several different approaches now being investigated for realization of this function. The approach considered in this paper is based on four-wave mixing using ultra-fast intraband dynamics in semiconductor traveling-wave amplifiers (TWA's). In what follows we will first quickly put wavelength converters into the context of all-optical networks and overview alternatives to four-wave mixing converters. We will then describe the physics of four-wave mixing in TWA's and show how four-wave mixing, in addition to its practical application to wavelength conversion, provides an important way of probing ultra-fast dynamics in TWA's. Finally, we will overview the practical issues associated with conversion of base-band digital information before concluding.

1.2 Wavelength conversion in all-optical global networks

Although consideration of all-optical global networks is still in the planning and test-bed phase, some design issues are by now fairly well established. Specifically, the design philosophy and architecture must produce a network which scales (both in geographic extent, information carrying capacity and number of users), accommodates many types of services, and minimizes cost by stressing system functions that are "transparent" to bit rate and also, where ever possible, to modulation format. A typical architecture is summarized in detail in reference [2] and will serve as the basis for this discussion. For our purposes here it is enough to consider the highly schematic version in figure 1.

In figure 1, three network layers appear and communication within the overall network utilizes wavelength division multiplexing. At the lowest level, a local area network (LAN) allows users to access the network through optical terminals. These LAN's would be administered by a single entity such as a corporation or university. Of interest here is that the architecture imposes "local" wavelengths that are not permitted to leave the LAN by use of wavelength-selective bypasses at the interface between the LAN and the next layer. This allows different LANs to reuse the same set of wavelengths thereby vastly increasing information carrying capacity.

Certain wavelengths are reserved for access to the intermediate layers which could be, for example, a metropolitan area network (MAN). The intermediate layer allows LANs to

communicate within a given MAN by use of wavelength routing devices. In addition, the intermediate layer provides an access to the top layer which would most likely be a fiber trunk-line.

Each layer would have electronic schedulers that coordinate with other schedulers in equivalent or other layers to work out "light paths" when a service is requested. The architecture would support point-to-point, point-to-multipoint (i.e., broad-cast), time-division multiplexed sessions (TDM), and a dedicated service for transmission of network control signals between the schedulers.

The need for wavelength conversion in this system occurs in the intermediate layer where it allows for improved flexibility in wavelength routing; however, this application is secondary in importance to its other application at the interface between the intermediate and top-layer. At this interface, conversion of information laden wavelengths in the intermediate layer to the available wavelengths on the trunk is absolutely essential for two reasons:

(1) To make possible efficient reuse of the LAN local wavelengths. These are available since the top layer is buffered by the intermediate layer from the LANs.

(2) To make global transmission scheduling possible using only local (i.e., point to point) information on available wavelengths.

The second of these is an extremely important feature of wavelength conversion at the top-layer interface. Without it, schedulers would need to consider vast amounts of information in determining a wavelength route. As the network expands, the computational complexity would become insurmountable.

All networks employing WDM will experience this problem at some size-scale, thus pointing out the very real need for wavelength conversion devices as a means to uncouple scheduling between layers and simplify scheduling algorithms.

1.3 Wavelength Conversion Techniques

There are many possible ways to accomplish the wavelength conversion function. The least sophisticated approach is signal detection and subsequent modulation of a laser at the new desired wavelength. This method becomes intractable as carrier wavelength count increases in a WDM network, nor does it satisfy the important requirement of bit-rate transparency and modulation format transparency. More sophisticated demonstrations to date of wavelength converters include using one of several available nonlinearities in semiconductor amplifier and laser structures or in silica fiber. All of these approaches fall into one of two categories: those which use cross-gain or phase saturation and those which employ some form of wave mixing.

In terms of technological maturity, the approaches based on cross gain or phase saturation are closest to practical realization. In these approaches an input signal saturates the optical gain of a

semiconductor laser (SL) or TWA and thereby changes the level of another signal that has been provided at the desired new wavelength (in the case of the SL this could be the lasing mode). Some of the earliest implementations of this idea were based on optical triggering induced by saturation in an SL [3]. More recently both SL's [4] and TWA's have been studied [5,6]. The non resonant (i.e., single pass) nature of TWA's gives them the added advantage over SL's of continuously tunable pump and signal waves. A further improvement along these lines has put the TWA into one arm of an monolithic interferometer and uses the attendant phase modulation associated with gain saturation to impose modulation on a new wave [7,8]. This approach allows for high contrast modulation at the new wavelength which can be a problem in the single-pass TWA cross-gain saturation conversion devices. It also has the advantage that input waves and converted output waves can be spatially separated by the interferometer. Finally, cross-gain and cross-phase saturation devices can be designed to be relatively polarization insensitive.

These advantages give these devices a practical edge for the time being over wave mixing based devices. However, all cross-gain or cross-phase modulation based techniques have an inherent limitation. In exchange for using the powerful interband nonlinearity associated with gain saturation, they are restricted to ASK modulation formats and single channel operation at modulation rates that are limited by the stimulated recombination time in these devices.

Wavelength converters based on four-wave mixing, as was also true with cross-gain and cross-phase saturation, can and has been done in both SL's [9] and TWA's, however, for similar reasons nearly all current work has focused on mixing in TWA structures. Mixing techniques have exploited all possible combinations of pump and signal mixing as illustrated in figure 2. This includes: mixing two cw pump waves in the active medium of a TWA and subsequent modulation of a signal wave to produce sidebands at the new desired wavelength(s) [10] (figure 2a); mixing an input signal with a first cw pump and subsequent modulation of a second pump to produce converted signal sidebands [11] (figure 2b); and finally mixing an input signal with a cw pump wave and subsequent modulation of this same pump wave to produce a new converted wave (figure 2c) which is phase conjugate to the original wave.

The last technique has attracted considerable attention owing to its simplicity in comparison to the other techniques (one pump wave vs. two) and also because it provides a converted signal which is the phase-conjugate replica of the input signal. This latter property, distinct from the wavelength conversion function has been used to compensate both fiber dispersion and fiber nonlinearities in transmission experiments [12, 13] as was first pointed out by Yariv and Pepper [14]. We will focus on the approach described in figure 2c throughout the remainder of this paper.

Four wave mixing wavelength conversion has been demonstrated in both silica fiber [15] and in semiconductor gain media [16,17,18]. In fiber, owing to the weakness of the intrinsic nonlinearity, very long (5-10 km) fiber lengths are required to achieve conversion efficiencies

around one percent (we define conversion efficiency as the ratio of converted wave power to input signal power). The long lengths required mean that phase mismatch is an important consideration in device design. In particular, operation near the zero dispersion point is required, thereby imposing a severe limitation on tunability. We also note that difference frequency generation using periodically domain inverted lithium niobate has also been used as a conversion technique [19].

Four-wave mixing in semiconductor TWA's, on the other hand, uses a series of ultra-fast and strong nonlinearities associated with intraband relaxation within the semiconductor. As will be shown later, the strength of these nonlinearities combined with the intrinsic optical gain of the device makes possible efficient conversion using short interaction lengths (typically about 1 mm). Phase mismatch is therefore not a serious consideration in these devices. In addition, compact and potentially monolithic converters are feasible.

As described earlier, four-wave mixing converters provide perfect contrast signal conversion that is transparent to bit rate and modulation format. It is the only conversion technique that can make these claims. However, at the present time most four-wave mixing based approaches require polarization control of the input signal wave to provide efficient mixing with the pump waves. Although certain schemes based on application of dual pump waves can allow for polarization independent operation, this issue remains a serious disadvantage. Likewise, separation of pump and signal waves is a critical issue with four-wave mixing techniques and although techniques to accomplish this exist and have been demonstrated [20], there is nothing yet as straightforward as the interferometric converters based on cross-phase saturation. Finally, as will be reviewed later, conversion efficiency and signal-to-noise in four-wave mixing based converters, while sufficient for systems experiments, at present remains lower than with techniques which employ cross-gain or cross-phase saturation. Nonetheless there is intense interest in these devices since they are new and offer unique and useful features for all-optical networks.

We now review the physics associated with the ultra-fast four-wave mixing dynamics in semiconductor TWA's. Apart from its importance to wavelength conversion, four-wave mixing in these devices has provided an entirely new method to probe ultra fast carrier dynamics. We will review some of the new results that have come from this work in recent years and also describe some new areas where four-wave mixing may provide additional insights into device physics.

II. Intraband Dynamics and TeraHertz Spectroscopy

2.1 Modeling Mixing Dynamics

The idea of using ultra fast intraband dynamics for four-wave mixing was first investigated theoretically by Agarwal [21]. This idea was not realized experimentally until the work of Tiemejer [22] and then later Kikuchiet. al. [23]. The use of the technique as a spectroscopic tool in the TeraHertz regime was done by Zhou, et. al [24,25]. Since that time there have appeared many experiemtal theoretical contributions that have provided increasing detail on additional intraband mechanisms. For a comprehensive theoretical overview of this subject, the reader is referred to references [26, 27, 28]. It should also be noted that the theory of ultra fast modal competition in semiconductor lasers explored originally by Bogatov is closely related to the present subject. Our purpose here is to provide a rapid overview of the essentials of four-wave mixing in a semiconductor TWA with sufficient detail to extract meaningful and useful results.

The configuration for four-wave mixing assumed here is illustrated in figure 3. For analytical simplicity, the pump wave is assumed to be much stronger then the input signal wave. Inside the TWA these waves mix, producing dynamic gain and index gratings which subsequently scatter energy from the pump and signal wave into new waves. We consider only scattering of the pump wave in this treatment since it is by assumption the strongest wave present in the guide. This scattering produces two new waves: one at the original input signal frequency which is related to the Bogatov coupling referred to above and the second wave, at a new shifted frequency, which is of interest in this paper. This second wave will be seen to be the phase conjugate replica of the original input signal field. As a result, it contains all of the original information contained in this wave. So, for example, if the input signal contains a base-band digital signal, then the new scattering wave will also contain this information. In this way, the four-wave mixing process can translate information from one region of the spectrum to another.

The mixing dynamics are describable using the coupled-mode equations for the complex field amplitudes. The equation of motion for the converted wave has the following form,

$$\frac{dE_1}{dz} = \imath k_1 E_1 + \imath\Gamma\frac{\omega_1}{2\mu c}\chi^{(3)}_{CD}E_2^2E_3^* \qquad (1)$$

where we have adopted a highly abbreviated form for the third order susceptibility, labeling it, for now, only by the subscript "CD" for carrier density modulation. In addition, we have absorbed the TWA optical gain into the wave vector, making it complex. Other quantities are defined in Table I. Similar equations hold for the other fields, however, the third order susceptibility is of

considerably less importance in these other equations under conditions where the undepleted pump approximation holds true. Also, note that the term involving the third order susceptibility contains the square of the pump field and the phase conjugate of the input signal field.

As noted earlier, phase mismatch is much less important in TWA converters than in fiber converters. By inspection of eqn. (1) it can be seen that phase mismatch considerations will require that,

$$(2k_2^r - k_1^r - k_3^r)\,L < 2\pi \qquad (2)$$

where we have ignored, for the moment, the effects of amplification (i.e., we treat the wave guide as transparent) by taking the propagation constants to be real. This expression can be shown to be equivalent to [26]:

$$\left|\frac{\partial \mu_g}{\partial \lambda}\right|\left(\frac{\Delta\omega}{\omega_2}\right)^2 L < 1 \qquad (3)$$

by using values that are typical for semiconductor TWA wave guides and taking an interaction length of 1 mm, we arrive at an allowable detuning frequency of about 8 THz. If we had properly accounted for amplification, this figure would be even larger since phase mismatch is frustrated to a certain extent by amplification along the interaction length (i.e., the effective interaction length is shorter than the actual device length).

The magnitude of the third-order susceptibility in eqn (1) is of central importance to four-wave mixing and we now investigate some of the mechanisms which contribute to this term in TWA's. For analytical simplicity we will restrict our attention to only two mechanisms: interband carrier density modulation and intraband spectral hole burning. In terms of dynamics, these respectively represent the slowest and the one of the fastest mixing mechanisms in a TWA. Our approach will bypass many important details so as to emphasize the essential physics. The results, however, will prove useful for estimating the strength of the mechanisms and in illustrating how other mechanisms can be incorporated into the coupled-mode equations phenomenologically.

Consider first carrier density modulation caused by the mixing of two waves at a point along a traveling-wave amplifier. Carrier modulation strength is determined by the dynamic balance between the rate at which carriers are added to or removed from the active medium as a result of stimulated emission and absorption (or taken together, the net gain).and the rate at which carriers relax to a local equilibrium density once perturbed. Linearizing the carrier density rate equation yields the following expression for the complex amplitude associated with carrier density modulation at the detuning frequency,

$$\delta\hat{n} = -g\frac{\varepsilon\,\tau_R}{h\,\omega}\frac{1}{1+\iota\,\tau_R\,\Delta\omega}E_2\,E_3^* \qquad (4)$$

where the relaxation time constant includes a contribution from stimulated emission. If, for the moment, we consider the susceptibility function to be a dynamic function of only carrier density, then the variation in the susceptibility caused by the time variation in carrier density described above will be given by,

$$\chi(n) = \chi_o + \chi_n\,[\delta\hat{n} + \delta\hat{n}^*] \qquad (5)$$

Upon substitution of eqn. (4) into this expression, we can immediately extract the third order susceptibility defined in eqn. (1). (In addition, although not done here, the third order susceptibility term arising from the conjugate carrier density amplitude in (5) can be seen to be the term associated with the Bogatov effect.) We give below the third-order susceptibility associated with carrier density modulation,

$$\chi_{CD}^{(3)} = -\chi_n g\frac{\varepsilon\,\tau_R}{h\,\omega}\frac{1}{1+\iota\,\tau_R\,\Delta\omega} \qquad (6)$$

This can also be written using the definition of differential gain and the alpha parameter as follows:

$$\chi_{CD}^{(3)} = \frac{\mu^2}{\omega}g_n(\iota+\alpha)g\frac{\varepsilon\,\tau_R}{h\,\omega}\frac{1}{1+\iota\,\tau_R\,\Delta\omega} \qquad (7)$$

In either of these expressions it is clear that a third order contribution to the susceptibility emerges as a result of harmonic saturation of the carrier density caused by the mixing of the input fields. The result also shows that this contribution depends strongly on the detuning frequency of the input signal and pump wave. Specifically, the corner frequency for this mixing process is set by the relaxation rate of the carrier enhanced by the stimulated emission. Typically, the time constant will be in the range of 200 psec in TWA's and therefore leads to a corner frequency in the range of several GHz.

Now consider including another mechanism, specifically spectral hole burning. Like carrier density pulsations, spectral hole burning results from stimulated emission induced saturation, however in this case the saturation is spectrally localized to those states that are resonant with the mixing waves. Consequently, the quantity which saturates is not total carrier density but rather the occupancy of the resonant states. For the purposes of our analysis we make an approximation

which divides the states into this resonant set and the remaining non-resonant set. The resonant states will fall within a spectral width about the optical waves determined by the polarization dephasing rate for interband transitions in the semiconductor. By using the density matrix formalism and applying the definition of resonant and non resonant states, we can show that the volumetric density of resonant states obeys an equation similar in form to eqn. (4). The approximation involves taking a partial summation of state occupancy over only the resonant states. The result is:

$$\delta\hat{n}_R = -g\frac{\varepsilon\,\tau_1}{h\,\omega}\frac{1}{1+\imath\,\tau_1\,\Delta\omega}E_2\,E_3^* \qquad (8)$$

where the stimulated decay time constant has been replaced by the state occupancy relaxation time constant. This time constant determines the rate at which state occupancy returns to equilibrium, once perturbed. It is very short (typically less than 100 fsec) and sets the practical rate for four-wave mixing by spectral hole burning.

To relate the variation in eqn. (8) to the susceptibility and in turn to eqn. (1), we must first consider some background on time scales in semiconductors and in turn their effect on how we define the susceptibility function. Figure 4 gives a simplified overview of relaxation within a given band of a semiconductor assuming that an impulse of occupancy is generated having an arbitrary energy distribution at time t=0. The fastest process to occur is describable by a relaxation rate which is not illustrated here, i.e., the dephasing rate of these states with other states in the same band or in other bands as a result of various scattering mechanisms. The next fastest rate of interest characterizes the transition of the distribution from part (a) to part (b) of the figure. This is the thermal equilibration rate of state occupancy within the band caused by Coulomb scattering and is assigned the time constant T1 just introduced in eqn. (8). At the conclusion of this process, the system is describable in terms of a Fermi distribution function having a quasi temperature and quasi Fermi energy. Next, in figure 4c, is energy relaxation by way of emission or absorption of phonons to the lattice temperature, and finally in 4d is the relaxation of the Fermi energy (equivalently the carrier density) to an equilibrium value by way of recombination with holes or electrons as the case may be.

In this progression, the susceptibility function can be expressed in ever coarser detail. In particular, by applying the density matrix formalism and rate equation approximation, we progress from a description in terms of the individual occupancies of all states, then to one based on Fermi functions involving the dynamical temperature and Fermi energy, and then to a yet simpler version in which electronic temperature is eliminated in favor of the lattice temperature leaving only the

carrier density (or equivalently quasi Fermi energy) as a dynamic variable. The time scales associated with these mechanisms are also shown in the figure 4.

We note before proceeding, that in addition to the rate constant model itself being an important simplification used here to characterize the system, we have tacitly assumed a single rate constant for each of these mechanisms which is in general not true as energy or crystal momentum dependence may appear in these quantities.

With this picture in mind we note that depending on the time scale, or equivalently the modulation frequency, under consideration it may be necessary to account for several or many dynamical variables in the argument of the susceptibility function. However, since the variations to state occupancy caused by wave mixing are themselves a perturbation to some well defined steady state, we can account for variations in susceptibility to first order by Taylor expansion about the steady state. Consequently, we can include carrier density modulation and occupancy modulation into a single equation as follows:

$$\chi = \chi_o + \chi_n [\delta\hat{n} + \delta\hat{n}^*] + \chi_{n_R} [\delta\hat{n}_R + \delta\hat{n}_R^*] \qquad (9)$$

It is important to remember that the last Taylor coefficient here results from perturbing only the resonant carrier density. One result of this fact is that the Taylor coefficient will be nearly a pure imaginary term proportional to the local change in gain. Proceeding as before in the case of carrier density modulation, we can now extract a spectral hole contribution to the third order susceptibility.

$$\chi_{SH}^{(3)} = -\chi_{n_R} g \frac{\varepsilon \tau_1}{\hbar \omega} \frac{1}{1 + \iota \tau_1 \Delta\omega} \qquad (10)$$

From eqn. (9) it is clear that this contribution is additive to the contribution of the carrier density modulation derived above. In addition, it will be weaker than the contribution from carrier density modulation (roughly in proportion to the ratio of the lifetimes), however, it will also have a much wider bandwidth response (assuming a 50 fsec. relaxation time constant, we would expect a corner frequency of 3.2 THz or equivalently a wavelength conversion span of 50 nm).

We can use these expressions to find an expression for the nonlinear refractive index. For example:

$$\mu_2^{CD} = \frac{\mu^2}{2c\omega} g_n (\iota + \alpha) g \frac{\tau_R}{\hbar \omega} \frac{1}{1 + \iota \tau_R \Delta\omega} \qquad (11)$$

where the units are inverse intensity.

By adding greater complexity to the model, it is possible to account for other important contributions to the four-wave mixing process including but not limited to a contribution from dynamic carrier heating associated with plasma absorption and stimulated emission. This has been done elsewhere.

For comparison with experiment, a multitude of mechanisms can be accounted for by use of the following phenomenological expression in which contributions associated with a specific mechanism are described in terms of a complex coupling constant and a relaxation lifetime [25]:

$$\chi^{(3)} = \sum_{m} \frac{c_m}{1 + \imath\, \tau_m \Delta\omega} \qquad (12)$$

2.2 TeraHertz Four-Wave Mixing Spectroscopy

By measuring the power versus detuning frequency of the converted signal produced by four-wave mixing in a TWA and properly accounting for output variations in pump and input signal power as they are tuned, a frequency response function (actually two different functions depending on whether detuning is positive or negative) results. This response function can be used to analyze ultra fast dynamics in the TWA by comparing it to the square magnitude of the function in (12). To date several groups have used this technique to infer information on spectral hole burning, dynamic carrier heating and other effects in semiconductor gain media [23, 24, 25, 29, 30].

Figure 5 shows an experimental setup that has been used by Zhou et. al [23, 24] to perform such measurements with extreme sensitivity. Input signal and pump waves are provided by high stability and narrow line width tunable erbium fiber ring lasers. These waves are injected into a

TWA at a fixed and controllable polarization. At the output appears four waves: the original input waves and two new waves resulting from the four-wave mixing process as illustrated in figure 3. The wavelength and the output power of the original waves is measured at the spectrometer. The new waves are measured in one of two ways. If strong enough they can be detected directly on the same spectrometer used to measure the original pump waves. If they are extremely weak, then a third ring laser can be used as an optical local oscillator to heterodyne detect the new waves using a pin diode and electrical spectrum analyzer (also shown in the figure). Although heterodyne detection provides the highest sensitivity for this process, in recent years it has proven possible to directly detect the new waves by using techniques that greatly improve conversion efficiency in the TWA. These techniques will be described in the next section.

In any case, the results of measurements that first appeared in reference [25] are shown in figure 6. Positive and negative detuned spectra for a tensile-strained multi-quantum-well amplifier. The maximum detuning frequency shown in the data is 1.7 THz. This corresponds to an equivalent temporal resolution of 92 fsec. Since the time that these data were taken higher detuning frequencies and correspondingly higher temporal resolutions have been obtained experimentally. Also shown in these figures are curves that result from applying the multi-time constant response model described in the previous section. For the purposes of comparison, figure 6 shows curves that result from applying a one, a two, and a three time constant fit to the data. A successful fit is only possible when at least three time constants are used in the model. The time constants and complex coupling coefficients used in these fits are given in reference [25]. It is important to note that a single set of constants is used to model both the positive and negative sideband spectra (i.e., all differences in the model result from a change in the sign of f).

The longest time constant is found to be 200 psec. and corresponds to the interband recombination rate enhanced by way of stimulated emission. The intermediate time constant is 650 fsec. and is believed to be associated with dynamic carrier heating, a mechanism that has been studied extensively using femtosecond pump-probe techniques on TWA's [31,32]. Finally, the third time constant is not fully resolved by this measurement and has an upper bound of 100 fsec. It is most likely the T1 relaxation time constant, although the contribution here could come from either spectral hole burning as described earlier or from another mechanism sometimes referred to as the delay in carrier heating [31, 32].

It is interesting to note that the time constants resulting from fitting of data taken using a compressively strained quantum well TWA are in close agreement with the above values for the tensile strained amplifier [25]. However, there are distinct differences in the complex coupling coefficients associated with the various four-wave mixing terms for the two cases. In addition, more significant differences have been noted in four-wave mixing spectra of bulk TWA's versus

quantum well TWA's. Many of these issues and differences will require further experiments to fully clarify.

One similarity that has been observed in all four-wave mixing spectra measured to date, whether taken using bulk or quantum well active layers, is a strong asymmetry in the positive and negative spectra. In particular, the mixing efficiency is always observed to be stronger for positive frequency detuning (equivalently negative wavelength downshifts for wavelength conversion). Modeling has shown that these differences arise from interferences which result between the various contributing four-wave mixing mechanisms. In particular, in regions of the detuning spectrum where two mechanisms become comparable in magnitude, interferences are possible. The phase of the coupling constants in devices measured to date is such that this interference tends to enhance the positively detuned frequency spectra and depress the negatively detuned spectra. These results are also predicted in density matrix analyses that have been done in the last few years.

Although THz four-wave mixing spectroscopy based on four-wave mixing is a excellent tool for study of ultra fast dynamics in the frequency domain, it is limited by the overlap of the various four-wave mixing mechanisms. As such, it is a tool that frequently must be used in conjunction with data from femtosecond pump probe experiments. In addition to the overlaps that cause the interferences noted above, the underlying dynamics responsible for a particular four-wave mixing mechanism are often more complicated that would be suggested by the simple multi-time constant picture presented above. (e.g., the delay in carrier heating mechanism -. for further discussion see reference [26,27]). Nonetheless the ability to view electronic occupancy response functions in the frequency domain up to and beyond THz rates provides an important qualitative difference in the study of ultra-fast carrier dynamics.

Before leaving this section we note that in addition to study of intraband dynamics, four-wave mixing spectroscopy has also been applied to probe transport in semiconductor quantum well systems at ultra-fast rates. In very recent work we have shown that by using samples containing two differing types of strained quantum wells (in our case alternating tensile and compressively strained quantum wells), it is possible to probe inter-quantum well transport [33]. In particular, the dipole matrix element for a direct optical transition in semiconductor quantum wells becomes strongly polarization dependent in the presence of strain. We have used this effect to induce localized photomixing in a quantum well having a particular strain type and to then probe an adjacent, opposing-strained well using an orthogonally-polarized third wave. The resulting spectra have so far been used to distinguish between the transport efficiencies of certain four-wave mixing mechanisms. Future work may make possible more precise determinations of inter-well equilibration rates.

III. Wavelength Conversion

3.1 Conversion efficiency and signal to noise

The two most important considerations in terms of the practical application of a four-wave mixing wavelength converter are conversion efficiency and the signal to noise of the converted wave. From the analysis presented in section II it is straightforward to show that the conversion efficiency is given by the following simple expression:

$$\eta = G^3 I_P^2 R(\Delta\lambda) \qquad (13)$$

where, in addition to the single pass gain and the input pump power, this includes a quantity referred to as the relative conversion efficiency function which contains all of the information on intraband dynamics responsible for the mixing process. There are several important observations to be made about the conversion efficiency function which have been noted by Zhou, et. al previously [18,34]. First, and foremost, the efficiency benefits from a numerically large nonlinearity (relative to other nonlinearities in other systems such as silica fiber) and hence relative conversion efficiency function. Second, the quadratic dependence on the input pump power for this process means that, in addition to large pump powers being desirable for high conversion efficiency, there is an optimal ratio between pump and input signal power for maximum converted power (this ratio being 2:1). Third, the cubic dependence of conversion efficiency on single pass saturated gain places a high premium on high gain TWA devices. The second and third points taken together say that to the list of TWA attributes for high four-wave conversion efficiency should be added large TWA saturation power. Essentially, the qualities which make for a good TWA also make for a good wavelength converter.

By mapping the wavelength shift dependence of the relative conversion efficiency function (essentially a repitition of the THz spectroscopy measurements) it is possible to estimate the required single pass gain for unity conversion efficiency at any desired wavelength shift. Figure 7 contains data on the dependence of R in a tensile strained amplifier over a wide range of up-conversion and down-conversion wavelength shift values. In Figure 8, this same data is used in conjunction with eqn. (15) to estimate the required single pass gain for unity conversion efficiency (at two input pump power levels). It is interesting to note that even for wavelength shifts as large as 50 nm the required single pass gain is still well within the realm of current TWA fabrication technology. It is also important to note that in performing this kind of analysis we have tacitly ignored any saturation dependence in the parameter R. Whereas R does depend on amplifier saturation or equivalently on the level of inversion, empirically we have found that in quantum well

devices it is relatively constant over a wide range of current and power levels. This lends some confidence to the previous assumption and to the predictions in figure 9. We do not expect that this empirical behavior will always hold true, however, and these results must therefore be viewed as estimates which are strictly true only when a device is not too deeply saturated.

Returning to eqn (13), it is clear that the optimization of conversion efficiency, is complicated by the coupling between single pass gain and the input pump power. The issues are further complicated by consideration of signal to noise in the process. The primary source of noise in a TWA converter is the introduction of amplified spontaneous emission (ASE) by the TWA in the spectral region into which the converted signal is generated. This noise level is given by the expression,

$$NP = 2\, n_{SP}\, (G - 1) h\omega B \qquad (14)$$

which points out that operation at high gain to improve overall efficiency comes at the expense of large quantities of ASE noise in the conversion band. Under these circumstances, overall optimization of signal-to-noise requires operation at large input pump power levels so as to saturate the TWA and thereby reduce ASE. Figure 9 shows data taken from ref [34] which illustrate this point. Converted power, ASE noise (into a 1 Angstrom bandwidth) and resulting signal to noise ratio are plotted versus total input power (signal and pump with optimal 2:1 ratio) for a 5 nm wavelength downshift in a tensile amplifier. The data clearly show steadily improving signal-to-noise levels with increasing input signal levels. As a result, conditions for optimum signal to noise ratio are not necessarily the same as for optimum conversion efficiency.

3.2 De-coupling conversion efficiency and signal-to-noise

The central problem in optimizing conversion efficiency and signal-to-noise in a TWA four-wave mixing converter is that these quantities are coupled. Recently, however, we have noted that this coupling can be eliminated almost entirely so that a device can be independently optimized with respect to both conversion efficiency and signal-to-noise [35]. This de-coupling rests on two simple observations: first, single pass gain can be transferred out of the region in which mixing occurs without loss of the useful cubic dependence noted in eqn. (15) provided that it occurs as preamplification, second, ASE noise generated by this preamplifier in the conversion band can be filtered between the preamplifier and the TWA which serves as the nonlinear element. In this scheme, the overall converter is made up of preamplifier, noise filter and mixer (i.e., deeply saturated TWA). The conversion efficiency and signal-to-noise equations are now given by:

$$NP = 2n_{SP}(g-1)h\omega B$$
$$\eta = G^3 I_P^2 R(\Delta\lambda) \qquad (15)$$

where G is the overall gain of the system (both preamplifier and mixer) and g is the residual gain of the mixer (necessary to maintain the nonlinearity and also to compensate for wave guide loss). The filter should ideally remove all ASE noise over a wide span of wavelengths in the conversion band.

As an illustration of this technique and a demonstration of wavelength conversion of base-band digital information, consider figure 10. Here we illustrate a simple optical link containing a converter of the type which separates mixer and preamplifier with a filter. In this case the filter is a simple fiber-notch filter having a bandwidth of 20 GHz, the preamplifier is an erbium fiber amplifier and the mixer is a tensile-strained multi-quantum well device. The pump laser is a tunable erbium fiber ring laser. Also included in the link are an optical receiver including several optical filters placed both before and after the receiver preamp to remove both ASE noise as well as to suppress the residual mixer pump wave and thereby prevent preamplifier saturation. The signal laser is a DFB laser having an 18 GHz direct modulation corner frequency. It was modulated at both 2.5 GB/s and 10 GB/s and the detected signal was analyzed using both a bit error rate tester and a microwave transition analyzer.

Figure 11a shows the spectrum of all signals at the output of the converter. The amplified input signal appears to the far right and is downshifted by approximately 8 nm. Also shown are the pump wave (center) and the converted signal wave. The conversion efficiency for this configuration is approximately 10%. To arrive at this number one must divide out the amplification factor present in the amplified input signal appearing in the display. Appearing immediately to the left of the converted signal is a spectral notch which shows the ASE filtering action provided by the notch filter. The signal-to-noise level in this case is 25 dB. In the actual measurement, the pump wave frequency is tuned slightly so as to bring the notch and the converted wave into coincidence. An eye diagram of converted data at 2.5 GB/s is shown in figure 11c and at 10 GB/s in 11d. The error rate in these measurement was 10(-10) and 10(-7), respectively, and is believed to be limited by the receiver preamplifier and not by the wavelength converter. Finally, in 11b is shown a converted data pattern of 11100100 at 10 GB/s.

IV. Conclusion

This paper has reviewed the need for devices called wavelength converters in all optical networks. Of the several competing techniques for realization of these devices, only four-wave mixing offers complete flexibility in terms of accommodating any desirable bit rate and modulation format (so-called optical transparency). After reviewing the physics of four-wave mixing as well as overviewing the application of four-wave techniques to study of THz dynamics in semiconductor gain media, we have considered wavelength conversion efficiency and signal to noise and their optimization. Results from a simple system demonstration at 2.5 GB/s and 10 GB/s were also presented.

References

[1] A. A. M. Saleh, OFC'92, paper ThC1.
[2]S. B. Alexander, J. Light. Tech., **11**, 714 (1993).
[3] H. Kawaguchi, K. Magari, H. Yasaki, M. Fukada, K. Oe, IEEE J. Quant. Electron., **QE-24,** 2153 (1988).
[4] T. Durhuus, R J S Pedersen, B. Mikkelsen, K E Stubkjaer, M. Oberg, S. Nilsson, *IEEE Phot. Tech. Lett.*, **5**, 86 (1993).
[5] B. Glance, J. M. Wiesenfeld, U. Kpren, A. H. Gnauck, H. M. Presby, and A. Jourdan, *Electron. Lett.* **28**, 1714 (1992).
[6] J. M. Wiesenfeld, B. Glance, J. S. Perino, A H Gnauck, IEEE Phot. Tech. Lett., **5,** 1300 (1993).
[7] B. Mikkelsen, T. Durhuus, C. Joergensen, RJS Pedersen, C. Braagaard, K E Stubkjaer, *Electron Lett.*, **30**, 260 (1994)
[8] M. Schilling, K. Daub, W. Idler, D. Baums, U. Koerner, E. Lach, G. Laube, K. Wunstel, *Electron. Lett.* **30**, 2128 (1994).
[9] S. Murata, A. Tomita, J. Shimizu, A. Suzuki, IEEE Phot. Tech. Lett., **3**, 1021 (1991).
[10] R. Schnabel, U. Hilbk, T. Hermes, P. Meissner, C. Helmolt, K. Magari, F. Raub, W. Pieper, F. J. Westphal, *IEEE Phot. Tech. Lett.*, **6**, 56 (1994).
[11] G. Grosskopf, R. Ludwig, H. G. Weber, Electron. Lett., **24**, 1106 (1988).
[12] M. C. Tatham, X. Gu, L. D. Westbrook, G. Sherlock, D. M. Spirit, *Electron. Lett.*, **30**, 1335 (1994).
[13] W. Pieper, C. Kurtzke, R. Schnabel, D. Breuer, R. Ludwig, K. Petermann, H. G. Weber, *Electron. Lett.* , **30**, 724 (1994).
[14] D. Pepper and A. Yariv, *Opt. Lett.*, **5**, 59 (1980).
[15] K. Inoue, and H. *Toba, IEEE Photon. Tech. Lett.* **4**, 69 (1992).
[16] M. C. Tatham, G. Sherlock, L. D. Westbrook, IEEE Phot. Tech. Lett., vol. 5, pp. 1303-1306 (1993).
[17] R. Ludwig, G. Raybon, *Electron. Lett.*, **30**, 338 (1994).
[18] J. Zhou, N. Park, J. W. Dawson, K. J. vahala, M. A. Newkirk, B. I. Miller, *IEEE Phot. Tech. Lett*, **6**, 50 (1994).
[19] C. Q. Xu, H. Okayama, K. Shinozaki, K. Watanabe, M. Karahara, *Appl. Phys. Lett.*, **63**, 3559 (1993).
[20] E A Swanson, J D Moores, *Photon. Tech. Lett.*, **6**, 1341 (1994).
[21] G. P. Agrawal, JOSA B, **5**, 147 (1988).
[22] L. F. Tiemeijer, *Appl. Phys. Lett.*, **59**, 499 (1991).
[23] K. Kikuchi, M. Kakui, C. E. Zah, T. P. Lee, *IEEE J. Quant. Electron.*, **28**, 151 (1992).
[24] J. Zhou, N. Park, J. W. Dawson, K. J. Vahala, M. A. Newkirk, B. I. Miller, *Appl. Phys. Lett.* **62**, 2301 (1993).
[25] J. Zhou, N. Park, J. W. Dawson, K. J. Vahala, M. A. Newkirk, B. I. Miller, Appl. Phys. Lett., **63**, pp. 1179-1181 (1993).
[26] A. Uskov, J. Mork, J. Mark, IEEE J. Quant. Electron., vol. 30, pp. 1769-1781 (1994).
[27] J. Mork and A. Mecozzi, Appl. Phys. Lett., Oct. issue (1994).
[28] A. Mecozzi, S. Scotti, A. D' Ottavi, E. Iannone, P. Spano, **QE-31**, 689 (1995).
[29] A. D'Ottavi, E. Iannone, A. Mecozzi, S. Scotti, P. Spano, J. Landreau, A. Ougazzaden, J. C. Bouley, *Appl. Phys. Lett.*, **64**, 2492 (1994).
[30] A. Uskov, J. Mork, J. Mark, M. C. Tatham, G. Sherlock, *Appl. Phys. Lett.*, **65**, 944 (1994).
[31] K. L. Hall, G. Lenz, E. P. Ippen, U. Koren, G. Raybon, *Appl. Phys Lett.*, **61**, 2512 (1992).
[32] J. Mork, M. Willatzen, J. Mark, M. Svendsen, C. P. Seltzer, **2146** SPIE Proceedings, Physics and Simulation of Optoelectronic Devices II, 24-26 January 1994, Los Angeles, Ca.

[33] J. Zhou, N. Park,, K. J. Vahala, M. A. Newkirk, B. I. Miller, *Appl. Phys. Lett.*, **65**, 1897 (1994).
[34] J. Zhou, N. Park,, K. J. Vahala, M. A. Newkirk, B. I. Miller, *IEEE Photon. Tech. Lett.*, **6**, 984 (1994).
[35] J. Zhou, K. J. Vahala, M. A. Newkirk, B. I. Miller, CLEO 1995, Paper CThT1.

TABLE I: DEFINITIONS

μ	Refractive Index
L	Amplifier length (interaction length)
$\Delta\omega$	Detuning frequency
$\delta\hat{n}$	Carrier density modulation amplitude
g	Optical Gain (Temporal rate units)
τ_R	Stimulated decay time constant
$\delta\hat{n}_R$	Resonant state modulation amplitude
τ_1	Occupancy equilibration time constant

Table I: Parameter and variable definitions.

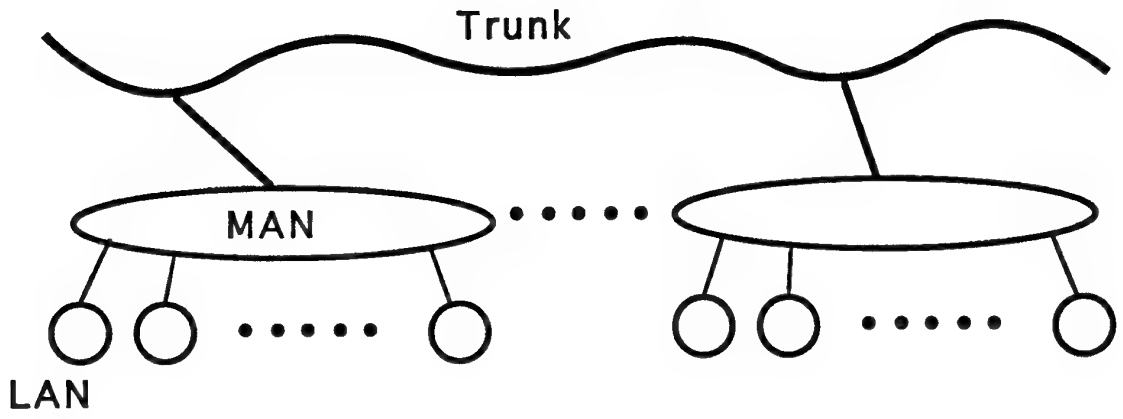


Figure 1: Highly simplified architecture for global optical network showing three layers. Wavelength converters improve wavelength routing flexibility in the intermediate layer and simplify scheduling access to the trunk.

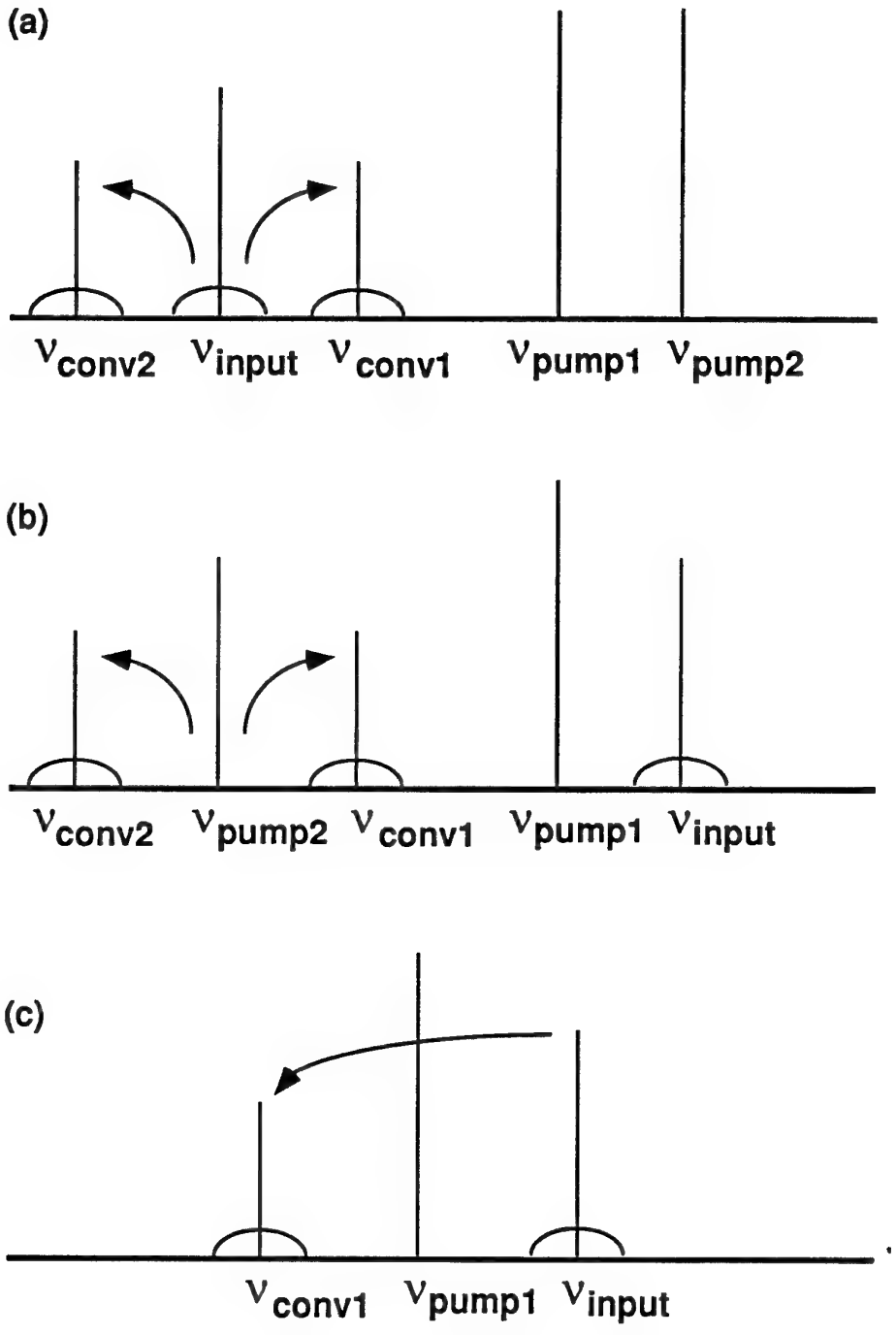
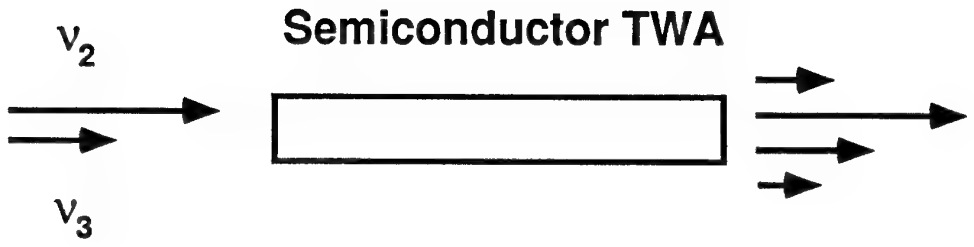


Figure 2: Various ways to apply four-wave mixing to achieve wavelength conversion.

FOUR-WAVE MIXING IN TWA'S



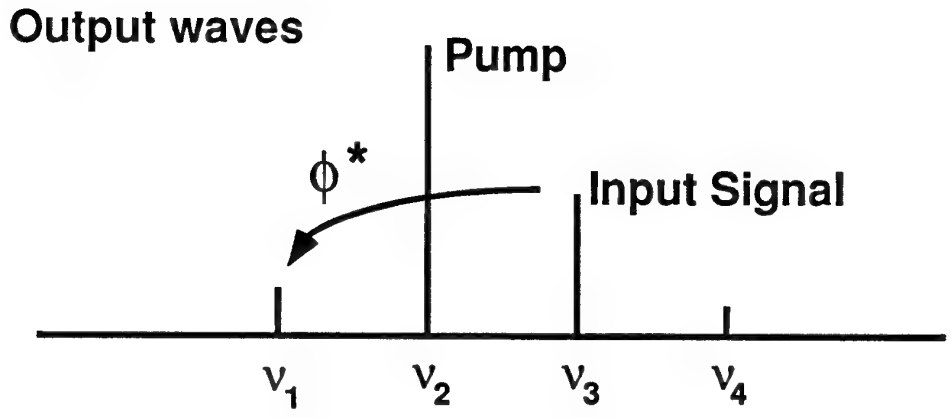
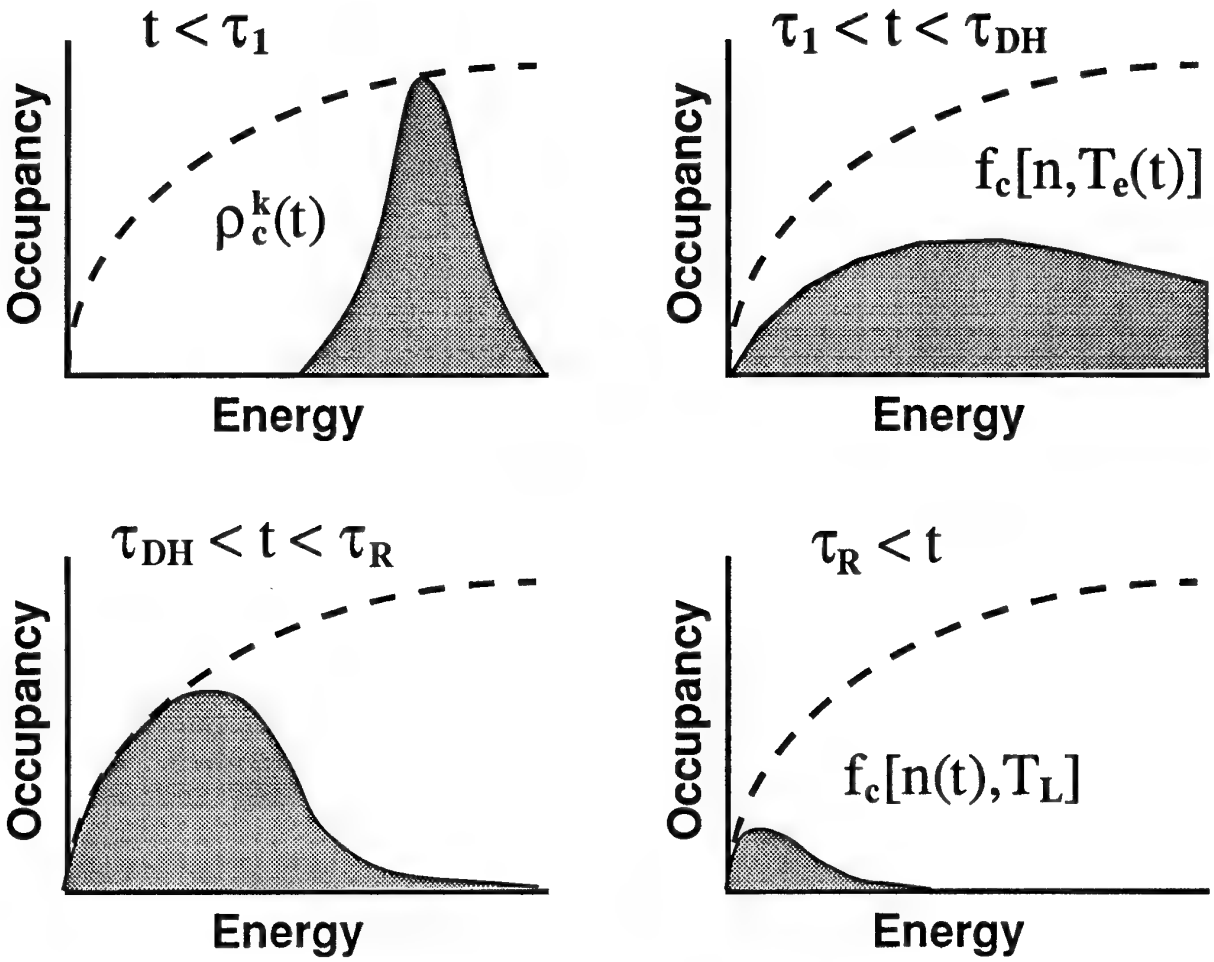


Figure 3: Typical experimental configuration for single-pump four-wave mixing.

Time-Scales for Population Relaxation



$$\tau_2 = \frac{1}{\gamma} \leq \tau_1$$

$$\tau_1 \approx 100 \text{ fsec.}$$

$$\tau_{DH} \approx 700 \text{ fsec.}$$

$$\tau_R \approx 200 \text{ psec.} - 1\text{nsec.}$$

2 to 3 Orders

Figure 4: Important time-scales in semiconductor gain media.

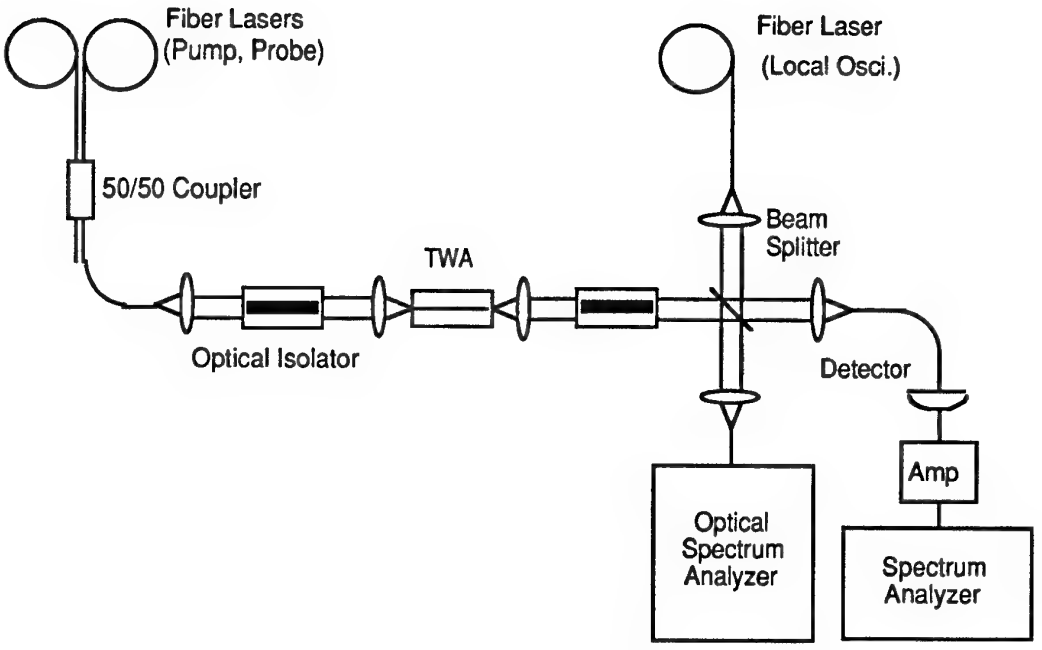


Figure 5: Experimental setup used in the four-wave mixing experiments.

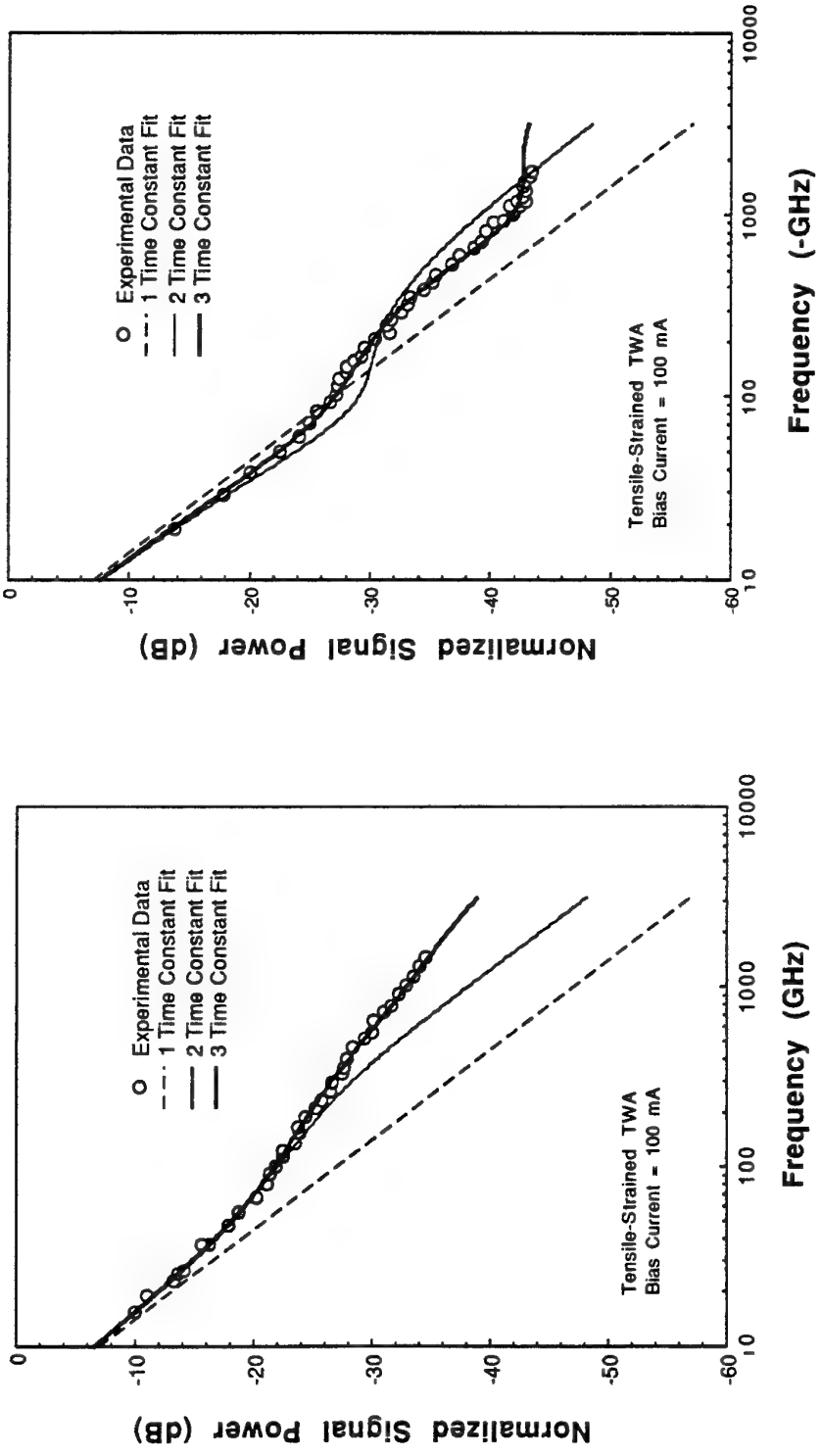


Figure 6: Positive and negative detuned four-wave spectra for a tensile-strained quantum well amplifier. Also shown is a one, two and three time-constant fit to the experimental data.

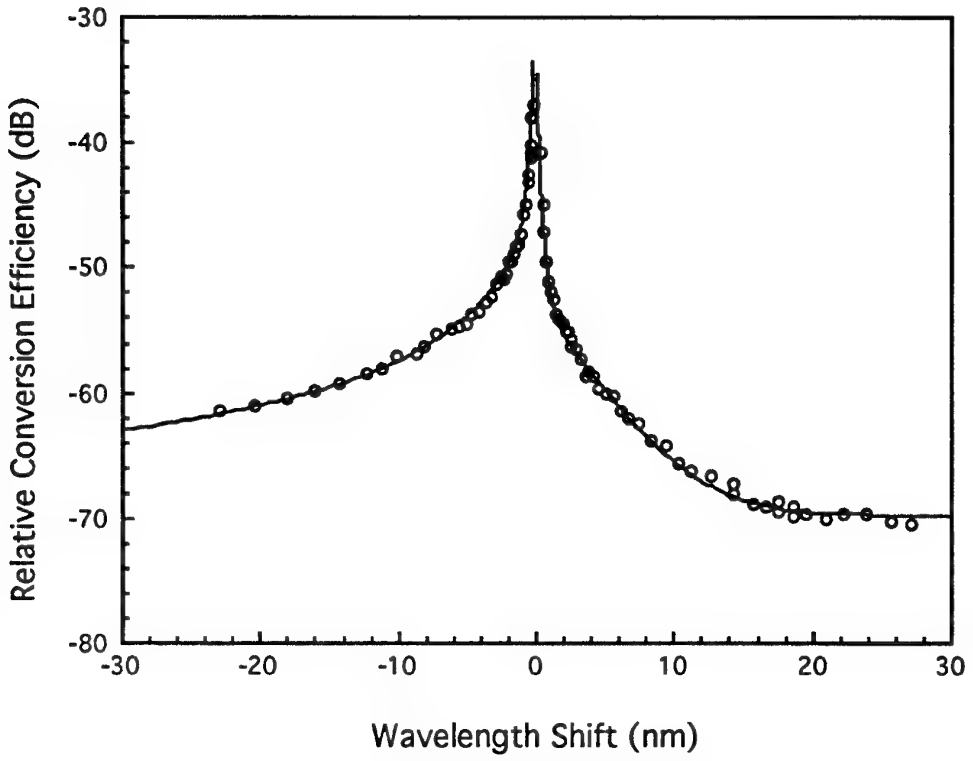


Figure 7: Measured relative efficiency function, $R(\Delta \lambda)$, versus wavelength shift.

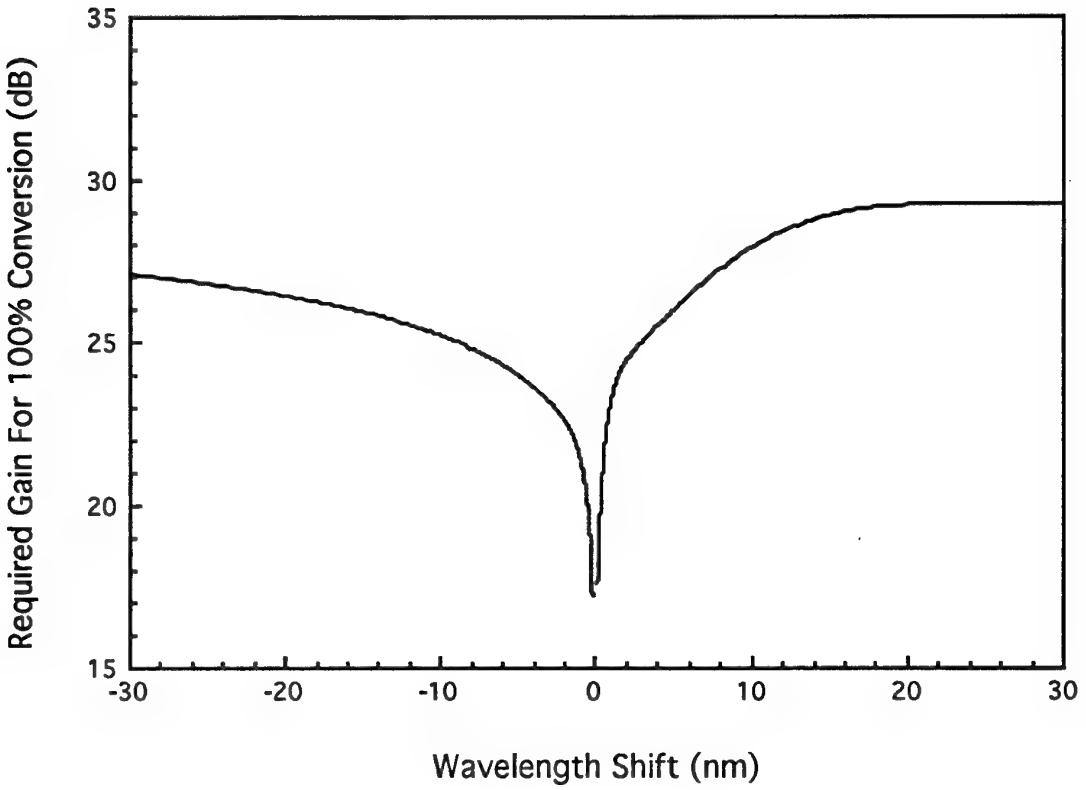


Figure 8: Single pass optical gain required for lossless wavelength conversion versus desired wavelength shift.

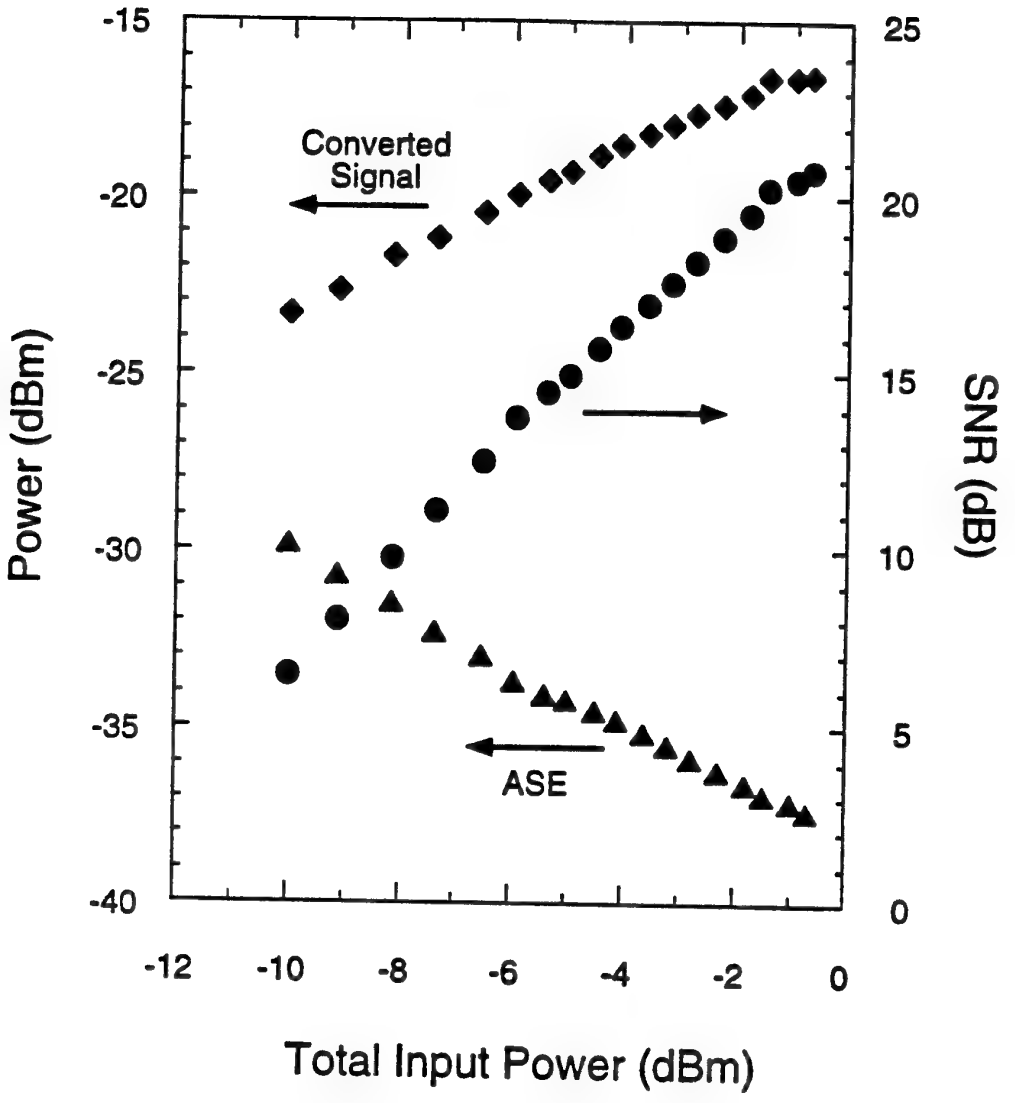


Figure 9: Converted power, noise power (into a 1 Angstrom bandwidth) and resulting signal to noise measured versus total input power.

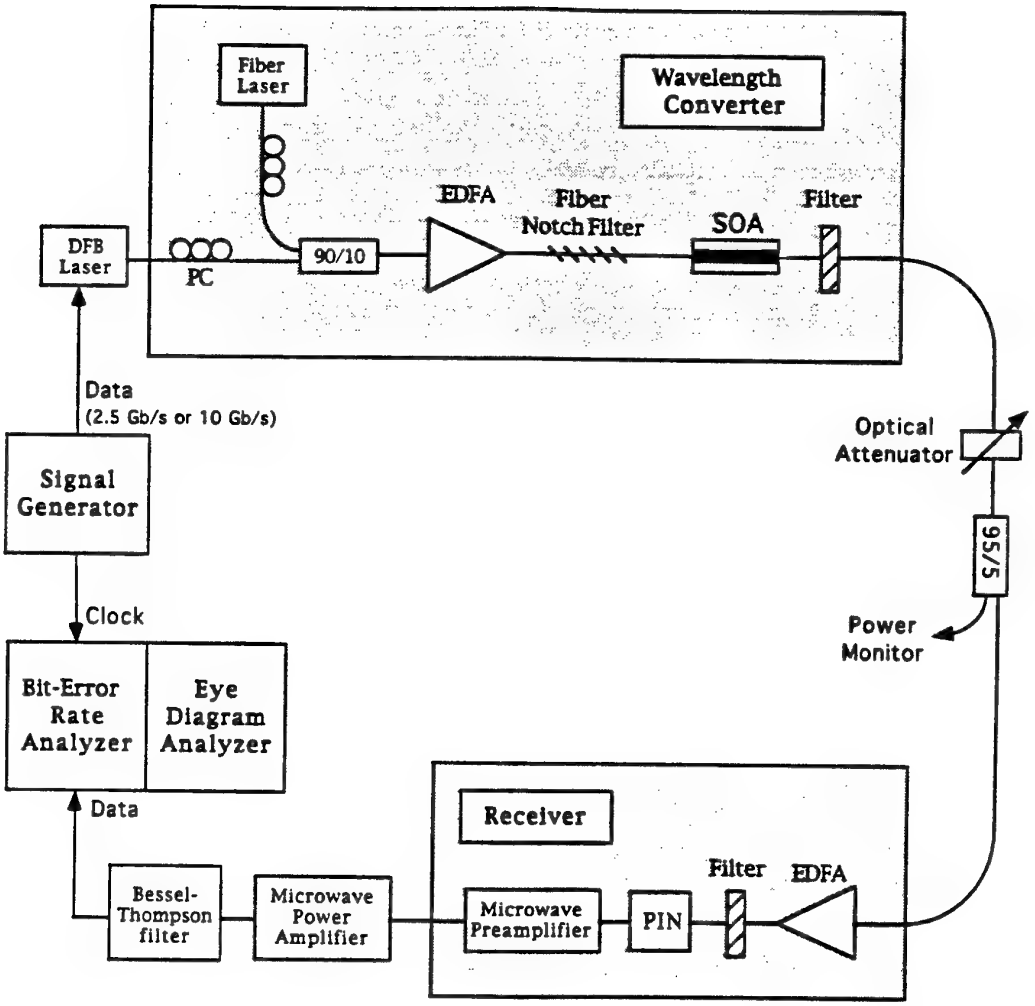


Figure 10: Experimental setup used for system test of four-wave mixing wavelength converter.

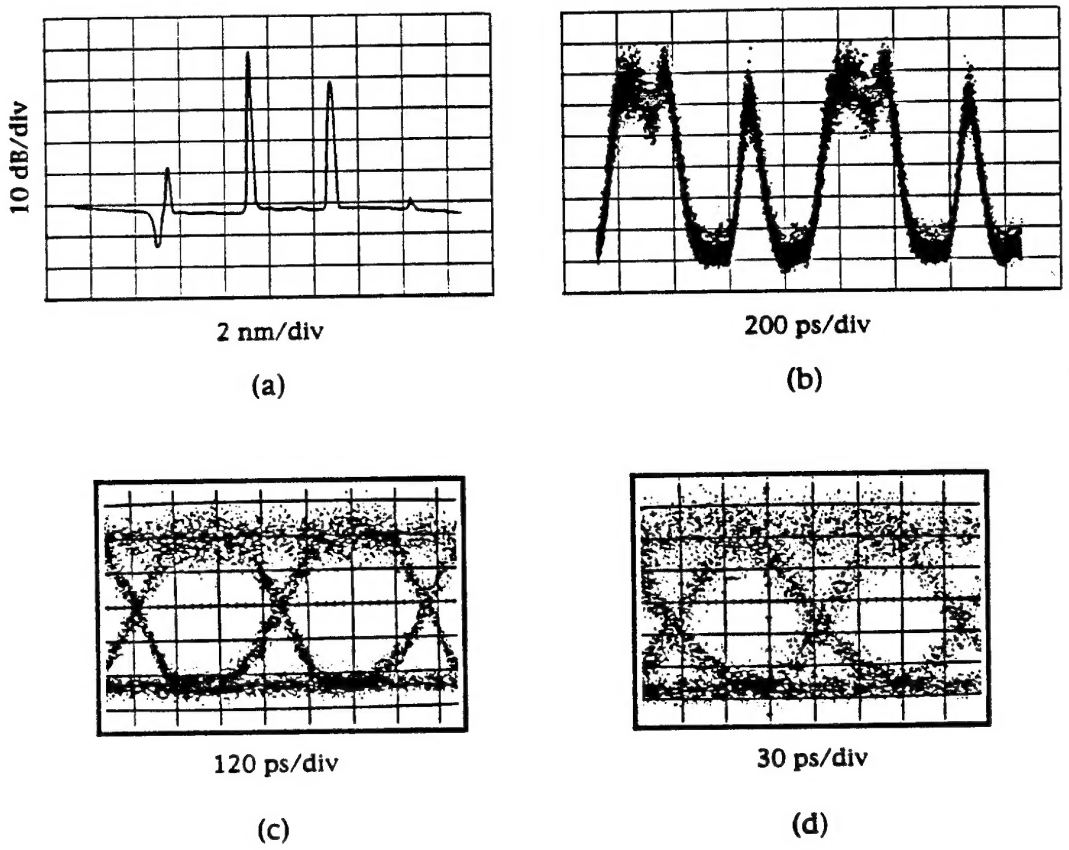


Figure 11 (a) Optical spectrum showing amplified input signal wave (far right), pump (middle) converted signal wave (left) and ASE notch. The resulting signal to noise is approximately 26 dB. (b) Converted data pattern of 11100100 at 10 GB/s. (c) Eye diagram for 2.5 GB/s converted data which has been wavelength shifted by about 8 nm. (d) Eye diagram for 10 GB/s converted data which has been wavelength shifted by about 8 nm.

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-LS-199	ISBN 92-836-0018-5	UNCLASSIFIED/ UNLIMITED

5. Originator Advisory Group for Aerospace Research and Development
North Atlantic Treaty Organization
7 rue Ancelle, 92200 Neuilly-sur-Seine, France

6. Title

Optical Processing and Computing

7. Presented at/sponsored by

The AGARD SPP Lectures Series held from 12-13 October 1995 in Paris, France, 16-17 October 1995 in Rome, Italy, and 19-20 October 1995 in Ankara, Turkey.

8. Author(s)/Editor(s)	9. Date
Multiple	September 1995
10. Author's/Editor's Address	**11. Pages**
Multiple	152

12. Distribution Statement There are no restrictions on the distribution of this document. Information about the availability of this and other AGARD unclassified publications is given on the back cover.

13. Keywords/Descriptors

Computer networks
Neural networks
Optics
Optoelectronics
Data processing
Signal processing
Optical processing
Optical storage
Optical computing

14. Abstract

Optical computing, namely information processing using light waves to represent the data, possesses some inherent advantage over electronic computing, in particular for massive data storage and parallel and neural processing. The main aim of the LS is to show how recent advances in lightwave technology make the time ripe to consider exploiting the potential of optical computing for data processing applications.

The LS will be opened with an overview of the basic concepts and inherent advantages of using optics for data processing and computing applications. The rest of the first day will be devoted to two topics: the use of optics for interconnecting electronic and optoelectronic processors and the use of optoelectronic techniques to enhance the performance of various computing devices and systems.

The second day of the LS will be opened with an overview of the emerging field of artificial neural networks as a signal processing paradigm, and its hardware, and in particular its optical implementations. Finally, the LS will be concluded with a description of recent developments of optoelectronic data communication, and their forecasted effect on computing and data processing technologies.

AGARD

NATO OTAN

7 RUE ANCELLE • 92200 NEUILLY-SUR-SEINE

FRANCE

Télécopie (1)47.38.57.99 • Télex 610 176

DIFFUSION DES PUBLICATIONS

AGARD NON CLASSIFIEES

Aucun stock de publications n'a existé à AGARD. A partir de 1993, AGARD détiendra un stock limité des publications associées aux cycles de conférences et cours spéciaux ainsi que les AGARDographies et les rapports des groupes de travail, organisés et publiés à partir de 1993 inclus. Les demandes de renseignements doivent être adressées à AGARD par lettre ou par fax à l'adresse indiquée ci-dessus. *Veuillez ne pas téléphoner.* La diffusion initiale de toutes les publications de l'AGARD est effectuée auprès des pays membres de l'OTAN par l'intermédiaire des centres de distribution nationaux indiqués ci-dessous. Des exemplaires supplémentaires peuvent parfois être obtenus auprès de ces centres (à l'exception des Etats-Unis). Si vous souhaitez recevoir toutes les publications de l'AGARD, ou simplement celles qui concernent certains Panels, vous pouvez demander à être inclu sur la liste d'envoi de l'un de ces centres. Les publications de l'AGARD sont en vente auprès des agences indiquées ci-dessous, sous forme de photocopie ou de microfiche.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Fachinformationszentrum,
Karlsruhe
D-76344 Eggenstein-Leopoldshafen 2

BELGIQUE
Coordonnateur AGARD-VSL
Etat-major de la Force aérienne
Quartier Reine Elisabeth
Rue d'Evere, 1140 Bruxelles

CANADA
Directeur, Services d'information scientifique
Ministère de la Défense nationale
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Establishment
Ryvangs Allé 1
P.O. Box 2715
DK-2100 Copenhagen Ø

ESPAGNE
INTA (AGARD Publications)
Pintor Rosales 34
28008 Madrid

ETATS-UNIS
NASA Headquarters
Code JOB-1
Washington, D.C. 20546

FRANCE
O.N.E.R.A. (Direction)
29, Avenue de la Division Leclerc
92322 Châtillon Cedex

GRECE
Hellenic Air Force
Air War College
Scientific and Technical Library
Dekelia Air Force Base
Dekelia, Athens TGA 1010

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
Aeroporto Pratica di Mare
00040 Pomezia (Roma)

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
N-2007 Kjeller

PAYS-BAS
Netherlands Delegation to AGARD
National Aerospace Laboratory NLR
P.O. Box 90502
1006 BM Amsterdam

PORTUGAL
Força Aérea Portuguesa
Centro de Documentação e Informação
Alfragide
2700 Amadora

ROYAUME-UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar-Ankara

Le centre de distribution national des Etats-Unis ne détient PAS de stocks des publications de l'AGARD.
D'éventuelles demandes de photocopies doivent être formulées directement auprès du NASA Center for AeroSpace Information (CASI) à l'adresse ci-dessous. Toute notification de changement d'adresse doit être fait également auprès de CASI.

AGENCES DE VENTE

NASA Center for
AeroSpace Information (CASI)
800 Elkridge Landing Road
Linthicum Heights, MD 21090-2934
Etats-Unis

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris
France

The British Library
Document Supply Division
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
Royaume-Uni

Les demandes de microfiches ou de photocopies de documents AGARD (y compris les demandes faites auprès du CASI) doivent comporter la dénomination AGARD, ainsi que le numéro de série d'AGARD (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Veuiller noter qu'il y a lieu de spécifier AGARD-R-nnn et AGARD-AR-nnn lors de la commande des rapports AGARD et des rapports consultatifs AGARD respectivement. Des références bibliographiques complètes ainsi que des résumés des publications AGARD figurent dans les journaux suivants:

Scientific and Technical Aerospace Reports (STAR)
publié par la NASA Scientific and Technical
Information Division
NASA Headquarters (JTT)
Washington D.C. 20546
Etats-Unis

Government Reports Announcements and Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 22161
Etats-Unis
(accessible également en mode interactif dans la base de
données bibliographiques en ligne du NTIS, et sur CD-ROM)

Imprimé par le Groupe Communication Canada
45, boul. Sacré-Cœur, Hull (Québec), Canada K1A 0S7

AGARD

NATO OTAN

7 RUE ANCELLE • 92200 NEUILLY-SUR-SEINE

FRANCE

Telefax (1)47.38.57.99 • Telex 610 176

DISTRIBUTION OF UNCLASSIFIED

AGARD PUBLICATIONS

AGARD holds limited quantities of the publications that accompanied Lecture Series and Special Courses held in 1993 or later, and of AGARDographs and Working Group reports published from 1993 onward. For details, write or send a telefax to the address given above. *Please do not telephone.*

AGARD does not hold stocks of publications that accompanied earlier Lecture Series or Courses or of any other publications. Initial distribution of all AGARD publications is made to NATO nations through the National Distribution Centres listed below. Further copies are sometimes available from these centres (except in the United States). If you have a need to receive all AGARD publications, or just those relating to one or more specific AGARD Panels, they may be willing to include you (or your organisation) on their distribution list. AGARD publications may be purchased from the Sales Agencies listed below, in photocopy or microfiche form.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordonnateur AGARD — VSL
Etat-major de la Force aérienne
Quartier Reine Elisabeth
Rue d'Evere, 1140 Bruxelles

CANADA
Director Scientific Information Services
Dept of National Defence
Ottawa, Ontario K1A 0K2

DENMARK
Danish Defence Research Establishment
Ryvangs Allé 1
P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (Direction)
29 Avenue de la Division Leclerc
92322 Châtillon Cedex

GERMANY
Fachinformationszentrum
Karlsruhe
D-76344 Eggenstein-Leopoldshafen 2

GREECE
Hellenic Air Force
Air War College
Scientific and Technical Library
Dekelia Air Force Base
Dekelia, Athens TGA 1010

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
Aeroporto Pratica di Mare
00040 Pomezia (Roma)

LUXEMBOURG
See Belgium

NETHERLANDS
Netherlands Delegation to AGARD
National Aerospace Laboratory, NLR
P.O. Box 90502
1006 BM Amsterdam

NORWAY
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
N-2007 Kjeller

PORTUGAL
Força Aérea Portuguesa
Centro de Documentação e Informação
Alfragide
2700 Amadora

SPAIN
INTA (AGARD Publications)
Pintor Rosales 34
28008 Madrid

TURKEY
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar-Ankara

UNITED KINGDOM
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES
NASA Headquarters
Code JOB-1
Washington, D.C. 20546

The United States National Distribution Centre does NOT hold stocks of AGARD publications.
Applications for copies should be made direct to the NASA Center for AeroSpace Information (CASI) at the address below. Change of address requests should also go to CASI.

SALES AGENCIES

NASA Center for
AeroSpace Information (CASI)
800 Elkridge Landing Road
Linthicum Heights, MD 21090-2934
United States

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris
France

The British Library
Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom

Requests for microfiches or photocopies of AGARD documents (including requests to CASI) should include the word 'AGARD' and the AGARD serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Note that AGARD Reports and Advisory Reports should be specified as AGARD-R-nnn and AGARD-AR-nnn, respectively. Full bibliographical references and abstracts of AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
published by NASA Scientific and Technical
Information Division
NASA Headquarters (JTT)
Washington D.C. 20546
United States

Government Reports Announcements and Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 22161
United States
(also available online in the NTIS Bibliographic
Database or on CD-ROM)

Printed by Canada Communication Group
45 Sacré-Cœur Blvd., Hull (Québec), Canada K1A 0S7

ISBN 92-836-0018-5